DOCUMENT RESUME

ED 209 343 TM 810 894

AUTHOR Haney, Walt; Gelberg, Wendy
TITLE Assessment in Early Childhood Education.
SPONS AGENCY Department of Education, Washington, D.C.
PUB DATE Dec 80
NOTE 148p.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Early Childhood Education; *Educational Assessment; Evaluation Methods; *Measurement Techniques; *Measures (Individuals); *Program Evaluation; Testing Problems
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

The goal of this booklet is to describe some of the special challenges posed by early childhood assessment in general, and particularly as they apply to Title I program evaluation. The booklet has four purposes: (1) to describe special issues in early childhood assessment; (2) to describe briefly alternative approaches to early childhood assessment; (3) to suggest how these issues relate to various purposes of assessment, particularly to that of Title I program evaluation; and (4) to provide some general guidelines on how to select and use early childhood tests and instruments. Subsequent chapters correspond to these four purposes, and appendices provide notes on recommended reading for further information on early childhood assessment, a listing of early childhood instruments and sources of review information on each, annotations to illustrate how potentially useful instruments can be initially screened, and descriptive reviews of instruments to illustrate information helpful in selecting among candidate instruments. The focus is on special issues in the educational assessment of young children. Educational assessment is broadly defined as systematic measurement, via testing or observation of individual behavior, traits, or other educationally relevant characteristics. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209343

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM PS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view. UD

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ASSESSMENT

IN

EARLY CHILDHOOD EDUCATION

December 1980

Walt Haney and Wendy Gelberg
The Huron Institute
123 Mt. Auburn Street
Cambridge, Massachusetts 02138

TM 800 894

FOREWORD

This booklet has been prepared as part of a project sponsored by the United States Education Department (USED) on evaluation in early childhood Title I (ECT-I) programs. It is one of a series of resource books developed in response to concerns expressed by state and local personnel about early childhood Title I programs. The series describes an array of diverse evaluation activities and outlines how each of these might contribute to improving local programs. The series revolves around a set of questions:

- Who will use the evaluation results?
- What kinds of information are users likely to find most helpful?
- In what ways might this information aid in program improvement?
- Are the potential benefits substantial enough to justify the cost and effort of evaluation?

Together, the resource books address a range of issues relevant to the evaluation of early childhood programs for educationally disadvantaged children. The series comprises the following volumes:

- Evaluating Title I Early Childhood Programs: An Overview
- Assessment in Early Childhood Education
- Short-Term Impact Evaluation of Early Childhood Title I Programs
- An Introduction to the Value-Added Model and Its Use in Short-Term Impact Assessment
- Evaluation Approaches: A Focus on Improving Early Childhood Title I Programs
- Longitudinal Evaluation Systems for Early Childhood Title I Programs
- Evaluating Title I Parent Education Programs

The development of this series follows extensive field work on ECT-I programs (Yurchak & Bryk, 1979). In the course of that research, we

identified a number of concerns that SEA and LEA officials had about ECT-I programs, and the kinds of information that might be helpful in addressing them. Each resource book in the series thus deals with a specific concern or set of concerns. The books and the evaluation approaches they describe do not, however, constitute a comprehensive evaluation system to be uniformly applied by all. Our feasibility analysis (Bryk, Apling, & Mathews, 1978) indicated that such a system could not efficiently respond to the specific issues of interest in any single district at any given time. Rather, LEA personnel might wish to draw upon one or more of the approaches we describe, tailoring their effort to fit the particular problem confronting them.

Finally, the resource books are not comprehensive technical manuals. Their purpose is to help local school personnel identify issues that might merit further examination and to guide the choice of suitable evaluation strategies to address those issues. Additional information and assistance in using the various evaluation strategies are available in the more technical publications cited at the end of each volume, and from the Technical Assistance Centers in the ten national regions.

TABLE OF CONTENTS

		Page
FOREWORD		i
I.	INTRODUCTION	1
II.	SPECIAL ISSUES IN EARLY CHILDHOOD ASSESSMENT	5
	Characteristics of Young Children that Make Assessment Difficult	
	Measurement Considerations	
III.	OBSERVATIONAL APPROACHES TO EARLY CHILDHOOD ASSESSMENT	19
	The Case for Observational Approaches to Early Childhood Assessment	
	Potentials and Limitations of Observational Approaches	
IV.	USES OF EARLY CHILDHOOD ASSESSMENT INSTRUMENTS	31
	Administrative and Public Accountability	
	Making Decisions Concerning Individual Students	
	Guidance to Teachers in the Classroom	
	Evaluation	
	Using Early Childhood Assessment for Multiple Purposes	
V.	SELECTING AND USING EARLY CHILDHOOD ASSESSMENT INSTRUMENTS	51
	Screening Potential Instruments	
	Trying the Test Out	
	Using and Interpreting Tests	
APPENDIX 1		61
	Notes on Sources of Further Information	
APPENDIX 2		73
	Listing of Selected Early Childhood Instruments and Sources of Review Information on Each	
APPENDIX 3		93
	Annotations on Early Childhood Instruments	
APPENDIX 4		99
	Reviews of Selected Early Childhood Instruments	
REFERENCES		143

I. INTRODUCTION

> Measurement of young children and their environments presents some special problems . . . because of the limited response system of the young child and the very rapid changes that occur early in life.
>
> S. Anderson et al., 1972

As this observation by a panel of experts in child development suggests, educational assessment of young children carries certain problems that do not necessarily arise in assessing older children. The goal of this booklet is to describe some of the special challenges posed by early childhood assessment in general, and particularly as they apply to Title I program evaluation. The booklet thus has four purposes:

- To describe special issues in early childhood assessment
- To describe briefly alternative approaches to early childhood assessment
- To suggest how these issues relate to various purposes of assessment, particularly to that of Title I program evaluation
- To provide some general guidelines on how to select and use early childhood tests and instruments.

Subsequent chapters correspond to these four purposes, and appendices provide:

- Notes on recommended reading for further information on early childhood assessment
- A listing of early childhood instruments and sources of review information on each
- Annotations to illustrate how potentially useful instruments can be initially screened
- Descriptive reviews of instruments to illustrate information helpful in selecting among candidate instruments.

Before going further, let us explain briefly why this booklet was written. It has been developed as part of a project, sponsored by the United States

Education Department, on evaluation of early childhood Title I programs. During an earlier stage of this project, state and local education personnel concerned with Title I expressed a variety of needs for information on early childhood testing and assessment (Bryk, Apling, & Mathews, 1978). In particular, these educators expressed:

- Frequent demands for information on technical and procedural problems in early childhood testing
- Concerns about the match between testing and early childhood program curricula
- Considerable interest in a wide range of tests and instruments, particularly ones concerning psychomotor, social, emotional and language development
- Interest in alternative means of assessment, including observation instruments and behavior inventories.

This booklet is a response to at least some of these needs. Its focus is on special issues in the educational assessment of young children. We define educational assessment broadly to mean systematic measurement, via testing or observation of individual behavior, traits, or other educationally relevant characteristics.* A narrower definition might be simply standardized testing. However, there are some very good reasons why early childhood assessment should not be confined to this form of measurement. We will elaborate on some of these reasons in Chapter 2. Here let us point out only that for many goals of early childhood education programs, no good paper-and-pencil tests are available. Hence, as many experts have pointed out (e.g. Walker, Bane & Bryk, 1973; Brooks & Weintraub, 1976; Goodwin & Driscoll, 1980), other forms of assessment--

---

* Some people define educational assessment even more broadly to include systematic measurement of characteristics and traits of educational programs and environments (e.g. Goodwin & Driscoll, 1980). For many purposes (for example evaluating program implementation), such assessment may be essential. However, in order to limit the scope of this resource book we focus mostly on assessment of children.

including systematic observation and rating scales--may be particularly appropriate for use with young children. For this reason, in Chapter 3, we will briefly review some of the potential benefits and drawbacks of alternative approaches to early childhood assessment.

II. SPECIAL ISSUES IN EARLY CHILDHOOD ASSESSMENT

What are the special issues in early childhood assessment that can cause problems? Why is it more difficult to assess young children than older children? There are two perspectives from which to answer these questions. The first deals with the nature of child development, and the characteristics of young children that make assessment difficult. The second treats these issues in terms of traditional measurement considerations: validity, reliability, and norms. The first two sections below describe these perspectives, and in the next chapter we describe some of the potentials and problems of observational and rating approaches to early childhood assessment.

CHARACTERISTICS OF YOUNG CHILDREN THAT MAKE ASSESSMENT DIFFICULT

As the quotation at the start of this booklet suggests, the assessment of young children is more difficult than that of older children. This is due not merely to measurement problems per se, but also to real and important features of how young children develop. Before discussing assessment issues from the measurement perspective, let us first summarize some of the features of child development that have implications for educational assessment.

One of the most obvious problems in the assessment of young children is that they cannot read and may lack other test-taking skills which we assume of older children. Thus, tests that require reading of instructions obviously cannot be used with young children. As an alternative, many tests for early elementary grades rely entirely upon oral instructions from the adult administering the test, and answer alternatives are presented in pictures or drawings. Yet even with oral instructions, children's short attention spans--at least with respect to tasks they have not chosen for themselves--may prevent them

from following directions correctly. Comprehending oral instructions, giving continued attention to a relevant item or picture, and marking or otherwise indicating a response alternative, all may be difficult tasks for young children, and may get in the way of assessing other skills or attributes of children. To cite one concrete example, young children may lack the fine motor skills necessary for marking some types of machine-scoreable answer sheets. For this reason, the use of separate answer sheets is generally not appropriate with early elementary children; and with preschool or kindergarten children, it may be necessary to use individual assessment procedures in which the test-giver marks the child's answer. For many young children this may be the only way to avoid confounding real skills of interest with clerical skills of testing taking. Also, when pictures or drawings are used in early childhood assessment, children may interpret them in unusual ways in light of their own experience.

A second issue which complicates assessment of young children is that their cognitive and affective development are not easily disentangled (Bradley & Caldwell, 1974). Cognition and affect seem to develop together in young children and to interact, making measurement of one dependent upon the other, until children are socialized into school and society and affective behavior becomes more stable. In other words, how children feel about a task or what mood they are in may easily influence their performance. Young children may have little interest in externally imposed tasks, and their attention to such tasks may easily wander (Pikunas, 1976). They may tire quickly (Illingsworth, 1972), and their responses to assessment procedures may be influenced by hunger, restlessness, desire to please, or a multitude of other motives and circumstances. The reactivity of young children thus makes their

performance in testing and assessment particularly susceptible to extraneous influences. Research suggests, for example, that young children's test performance is more apt to be influenced by situational variables--including the ethnicity of the test giver--than that of older children (Epps, 1974).

Children's interpretation of assessment tasks and questions may also depend on their level of development. Young children's tendency to view things in relation to themselves and their experience (often called children's egocentrism) may prevent them from interpreting a question in the way an adult expects. Two examples will help to illustrate this point. On a standardized test, when one young boy read a short reading passage and then was asked why, in a test item, a girl named Susan watches television, he marked the response alternative expected from the passage: "Because Susan likes to watch television." Yet when the child was asked to explain his answer, he said: "Because I like to watch television." His answer derived not from the story, but from his egocentric perspective of why he watches television. In another example from a first grade reading test, children were instructed to mark the one picture out of three that goes best with the word next to the pictures. One item contained the word "fly" with an arrow pointing to pictures of an elephant, a bird, and a dog. Instead of marking the intended answer, the bird, many first graders had chosen the elephant, or the bird and the elephant. Asked to explain their answers, children identified the elephant as "Dumbo," the flying elephant (Mehan, 1978, p.51). In short, whatever their chronological age or grade level, children's interpretations of assessment tasks may be strongly influenced by many factors, including their personal experience and how they feel at the time of assessment.

In their early years, children develop rapidly, and while all children tend to pass through the same stages of development, they may do so at different rates. These two aspects of child development greatly complicate the use of systematic procedures in early childhood assessment. A procedure that is appropriate for one five-year-old may work not at all for another. As one child development expert put it:

> While the developmental rate is high during the preschool years, great variability in scores from successive testings is not uncommon. An appreciable degree of consistency emerges only after about age five when the developmental rate has slowed greatly and when going to school brings a relatively common program of environmental encounters into the lives of children.
>
> (Hunt, 1961, p.313)

MEASUREMENT CONSIDERATIONS

The various characteristics of young children that make assessment and testing more difficult than that with older children can also be viewed from another perspective: that of the measurement qualities of assessment procedures. In particular, tests and instruments for use with young children are generally of lower technical quality than those for use with older children. Measurement experts have made this observation (e.g. Goodwin & Driscoll, 1980; Walker, Bane, & Bryk, 1973), and the point has also been shown in systematic reviews of tests. When the Center for the Study of Evaluation (CSE) of the University of California at Los Angeles reviewed some 800 published standardized tests for the elementary school level, including over 3,900 subtests, they found that only nine of the first grade level subtests--or less than one percent--received minimally satisfactory ratings.

At higher grade levels, both the numbers and proportions of tests CSE rated as minimally adequate increased steadily (Hoepfner et al., 1976). At the first grade level, only in the domain labeled "cognitive and intellectual skills" was more than a single test rated as minimally adequate.* Such evaluations confirm that early childhood tests lack the technical qualities of later-grade tests. This general contrast also tends to hold with respect to the specific measurement qualities of validity, reliability, and norms.

Validity

The most important aspect of assessment quality is validity--that is, whether an assessment instrument really does measure what it purports to measure. Though people often speak of validity as if it were a characteristic of a test or assessment instrument, this is not really appropriate. Strictly speaking, the validity of an assessment procedure resides not in the instrument itself, but in its use in a particular way with a particular population. As Cronbach observed, "one validates not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971, p.447). Exactly how to conduct such validation is still a point of considerable debate among measurement experts, but three types of validity criteria are widely recognized:

- Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent.

---

* The CSE also rated quality of prekindergarten and kindergarten tests in an earlier study (Hoepfner et al., 1971). However, since the rating scheme was slightly different in this study than in the one cited above on elementary level tests, the results cannot be directly compared. See Haney et al., 1978, pp.111-119, for details of ratings across grade levels and measurement domains as well as for a review of criticisms of the CSE approach to rating test quality.

- Construct validity is implied when one evaluates a test or other set of operations in light of a specified construct--that is, an idea developed or "constructed" as a work of informed, scientific imagination, such as "intelligence," "readiness," or "social competence." In other words, a construct is a theoretical idea developed to explain and organize some aspect of existing knowledge.

- Criterion-related validity applies when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. There are two forms of criterion-related validity. Predictive validity refers to inferences regarding future performance, while concurrent validity refers to inferences concerning performance observed or measured at approximately the same time as testing or assessment takes place.

(APA, AERA & NCME, 1974)

Although test publishers tend to emphasize content validity in documenting the quality of their instruments, some experts have argued strongly for the importance of construct validity instead of content or criterion-related validity (Messick, 1975). Also, others have recently argued that for educational purposes, tests should have curriculum and instructional validity, i.e., they should be related to the content of curriculum and instruction. Since such arguments cannot be resolved in the abstract, let us simply discuss some of the general validity considerations in early childhood assessment.

Two of the most common types of early childhood instruments are intelligence tests and readiness tests. Indeed, intelligence and readiness are two of the most familiar constructs in early childhood assessment. Yet in practice there is much confusion about what it is that each of these terms or constructs actually encompasses. For example, in their test evaluation project, CSE investigators actually classified subtests of some intelligence tests as measuring "readiness skills" (Haney et al., 1978, p. 120). In a recent federal court case in California, lengthy expert testimony revealed the widely conflicting opinion and confusion that exists in the field of

educational and psychological measurement over the meaning of intelligence and whether and how tests of intelligence relate to it. This confusion was one of the reasons why the judge in the case ruled that intelligence tests are biased against minority children and illegal to use in placing children in classes for the educable mentally retarded (Larry P. v. Riles, 1979).*

Similar confusion surrounds the term and construct of readiness. School and reading readiness tests are commonly used in early education programs in the United States and have been published in this country for at least half a century. Yet there is still little agreement about what constitutes school readiness or reading readiness. Reflecting this disagreement, reading readiness tests vary considerably in the skills they cover. This was shown by Rude (1973) in his analysis of five major reading readiness batteries to determine which of twelve specific skills were actually assessed. The number of skills assessed on any one readiness battery ranged from three to seven. Eight skills were assessed on only one of the batteries. The only skill that was assessed in all five batteries was letter recognition. The main problem with all such readiness tests is that one cannot gauge their value without specifically addressing the question of readiness for what: not just readiness for first grade or for reading, but for what kind of first grade or reading.

Disagreement and confusion over what is meant by reading readiness and intelligence does not mean that tests which go by these names are useless. It does mean, however, that if one wants to use an early childhood instrument for a particular purpose, one should not simply accept an instrument at face value and assume that it measures a construct such as intelligence or readiness, but should carefully examine the nature and validity

* The initial decision in the Larry P. case is currently under appeal.

of the instrument in light of that purpose. If an instrument is to be used to help select children for future participation in special programs like Title I, then attention should be given to its predictive validity--that is, to how well results will predict children's future performance. If the instrument is to be used to evaluate a program, then consideration should be given to how well the content of the instrument matches the content of the program of instruction. If one wishes to use an instrument to infer something about a construct or general aspect of children's development, say general reading achievement, then special consideration needs to be given to construct validity. In short, different potential uses of a test or other assessment device require attention to different kinds of validity evidence.

This point, which is relevant to testing and assessment generally, is especially important with respect to assessment of young children, since several extraneous aspects of assessment can have a strong influence on results for young children. Research suggests, for example, that how test instructions are given to children before assessment can affect results more for younger than for older children (Gaffney & Maguire, 1971). Also, use of separate machine-scoreable answer sheets can affect test performance of young children more than that of older children (Ramseyer & Cashen, 1971). Indeed, one test expert has advised: "In testing children below the fifth grade, the use of <u>any</u> separate answer sheet may significantly lower scores . . . . [At lower] grade levels, having the child mark the answers in the test booklet itself is generally preferable" (Anastasi, 1976, p.36). These issues in early childhood assessment are not, of course, simply assessment problems; rather, they reflect important characteristics of child development

discussed above. They can also affect the qualities of testing and assessment known as reliability and norming.

Reliability and Measurement Error

Reliability refers to the accuracy or consistency of measurement. Three types of reliability are most commonly treated in the educational measurement literature:

- Internal consistency refers to the extent to which all items or parts of an assessment measure the same thing
- Alternate form reliability means the comparative accuracy of results from equivalent forms of the same assessment instrument
- Stability refers to the consistency of assessment results over time.

(APA, AERA & NCME, 1974)

Although these three types of reliability have been widely recognized in the past, numerous sources of error in assessment are intertwined with far more complexity than is represented by just these three (Cronbach, Gleser, Nanda, and Rajaratnam, 1972). Indeed, when pursued thoroughly, issues of reliability or dependability begin to merge with issues of validity. And like validity, reliability cannot be treated very sensibly in the abstract.

Some people have tried to rate test reliability independently of test use,* but this ignores the obvious point that reliability of assessment is

---

* In rating elementary and school tests, for example, CSE investigations awarded three points to any test reporting an internal consistency coefficient greater than .90, two points if the coefficient "ranged from .70 to .90, one point if less than .70, and zero points if no appropriate coefficient was reported" (Hoepfner et al., 1976, p.xxix.) Points were awarded similarly for test stability and alternate form reliability.

more important for some uses than for others. In general, the more consequential is the assessment, the more we ought to be concerned with test reliability or accuracy. If a test is to be used to select children for a special program of some duration, such as Title I, then reliability matters far more than if it is to be used only for a monthly check on children's progress. Also, the intended use for a test or other assessment will affect the form of reliability evidence that should be considered. If a test is to be given in the spring as a means of helping to decide which children should receive Title I services in the fall, the stability of test scores over time would be an extremely important aspect of reliability. For other types of use, other aspects of reliability would be more pertinent.

Like validity, reliability of assessment can be more problematic with young children than with older ones. Indeed, it poses a special dilemma for early childhood assessment. Internal consistency and stability both tend to be lower in assessment of young children than in that of older ones (Walker, Bane & Bryk, 1973, p. 26; Brooks & Weintraub, 1976, p.39). As noted above, young children generally have shorter attention spans than older children--at least for tasks that are not of their own choosing. As a result it is important that assessment tasks for young children be kept short. The problem this raises, however, is that the shorter the test--that is, the fewer items it encompasses--the lower its reliability will be. Developers of early childhood tests and instruments get around this problem in several ways. First, they often organize assessment procedures into several relatively short sessions--of only 10 to 15 minutes for kindergarten-aged children and 15 to 20 minutes for first graders. This can help to avoid problems of inattention and fatigue that would likely result from longer

sessions. Second, many early childhood assessment instruments are individually rather than group administered--which can also help to maintain children's interest. Third, assessment tasks can be designed so as to be of intrinsic interest to children--indeed, some publishers of early childhood tests suggest that they should be described to children not as tests, but as games.

Norms

The third aspect of technical quality that should be mentioned is norms. Norms represent the performance on an instrument of some sample of persons with whom the instrument was standardized or normed. Norms are "empirically established by determining what a representative group of persons actually do on the test" (Anastasi, 1976, p.76). A score derived from the test or assessment procedure can then be interpreted in terms of the distribution of scores obtained by the group who participated in the instrument's norming or standardization.

For many early childhood tests and instruments, norms are nonexistent, or if available, are limited in certain respects. Early childhood instruments that are designed to assess children's performance on specific tasks, for example to help ascertain whether children can do certain things like tying their shoes or saying their names, may have no norms. When early childhood test norms do exist, they generally are not based on nationally representative samples.

When early childhood test norms are available, they may be limited in other respects. Some readiness tests have start-of-school-year norms, for example, but since they are designed as screening instruments to

assess children's status upon school entry, no empirical norms may be available for end-of-year performance. This constrast also reflects the point, noted above, that young children develop rapidly. A test which is useful with a group of five-year-olds or six-year-olds in the fall may simply not be useful with them the following spring. Also, children's performance on early childhood tests can be sharply influenced by pre-school or early school experience, which can complicate use and interpretation of norm-referenced results. The Comprehensive Test of Basic Skills (CTBS) Level A (Form S), for example, provides two sets of norms for the beginning of first grade--one for students who attended kindergarten and one for students who did not. On the alphabet subtest of the CTBS Level A, a particular raw score can vary by as much as 40 percentile points when interpreted in terms of the two sets of norms (CTB/McGraw Hill, 1974). In similar fashion, children's kindergarten performance can be sharply affected by whether or not they have attended prekindergarten. This complicates norm-referenced interpretations of early childhood test results, because "the experiences of the preschool child are less uniform than those of older children who are attending school" (Broman, Nichols & Kennedy, 1975, p.38).

In sum, early childhood tests and instruments tend to be of lower technical quality than those designed for use with older children. Validity of assessment of particular attributes of young children may be threatened when assessment results are confounded with aspects of assessment procedures such as children's skill in listening and ability to follow directions. Reliabilities of early childhood assessment instruments tend to be lower than those of instruments for older children. And norms for early childhood

instruments are often unavailable or are based on samples of children far smaller than those used in norming later grades' tests. Interpretation of norm-referenced results with younger children is also complicated by the fact that children's preschool or early school experience can sharply affect such results. The point that should be stressed, however, is that these qualities of early childhood tests and instruments do not reflect technical issues so much as they represent real and important characteristics of young children--that they grow and develop rapidly, that aspects of their cognitive, social, and affective development interact, and that they are not so accustomed to school and the procedures of educational assessment as are older children.

III. OBSERVATIONAL APPROACHES TO EARLY CHILDHOOD ASSESSMENT

Because of the many factors which can complicate the educational assessment of young children, alternative approaches to assessment, with alternative strengths and weaknesses, may be useful. These alternatives include interviews with children, documentation and recording of their educational activities and interests, and talking with parents about their children's learning. Such assessment techniques are, of course, nothing new. Teachers of young children typically rely upon just such varied assessments for a variety of purposes. Though most often used informally, such approaches can also be adapted to purposes of systematic assessment. Experience with large-scale evaluation has shown, for example, that techniques such as structured interviews with parents can illuminate aspects of early childhood programs which cannot be illuminated directly through traditional testing of children (see Haney & Pennington, 1978, for an example of how analyses of systematic parent interviews were used in this way with respect to Project Follow Through).

In this chapter we briefly describe several varieties of a general form of educational assessment which is often overlooked, namely systematic observation. First, we discuss why observation can be an especially useful approach to assessment of young children and describe exactly what is meant by systematic observation. Second, we briefly describe five types of systematic observation and an example of each. Third, we describe some of the general potential value and the limitations of observational approaches to early childhood assessment.

THE CASE FOR OBSERVATIONAL APPROACHES TO EARLY CHILDHOOD ASSESSMENT

Two experts recently summed up the case for using observational techniques in early childhood assessment as follows:

> Observational measurement is of particular importance in early childhood education for three reasons. First, and possibly most important, it affords a means of measuring many child behaviors that might otherwise be unmeasurable. Very young children, say five years and under, have a limited response repertoire, and especially if verbal-related. Thus, they may be unable to make the response or provide the information that a more conventional measure, such as an interview or a paper-and-pencil test, may require. Observational measurement may offer particular advantage in the affective domain. . . .
>
> A second reason for the appropriateness of observational measurement in early childhood education is that young children frequently fail to take testing procedures seriously. . . .
>
> The third reason relates to the generally held assumption that very young children are open and relatively unchanged or unperturbed by being observed.
>
> (Goodwin & Driscoll, 1980, p.111)

Before describing different types and examples of observational approaches to early childhood assessment, let us specifically explain what is meant by the term. Observational measurement refers to the systematic recording of the behavior or other characteristics of children. This includes use of checklists, rating scales, and observation scales and many individually administered early childhood tests in which the examiner rather than the child records children's responses to assessment tasks. Indeed, the fact that portions of commercially published early childhood "tests" such as the CIRCUS and the McCarthy Scales (both described in Appendix 4) call for the examiner's recording or rating of children's responses is clear testimony to the importance of not confounding the assessment of children's

characteristics or behaviors with their test-taking skills in general and their skill in recording answers in particular. This point cannot be overemphasized. Research has shown, for example, that scores on paper-and-pencil tests of children's "self-concept" may correlate more highly with children's performance on paper-and-pencil tests of achievement than they do with one another (see Haney, 1977, pp. 319-322, for a discussion of just such a pattern of results in the national Follow Through evaluation). In short, paper-and-pencil tests may confound young children's test-taking skills with other attributes they intend to measure. For such reasons, many experts (e.g. Walker, 1973, p. 38) have suggested that non-verbal observational techniques may be more valid and reliable means of measuring many characteristics of young children, particularly non-cognitive ones.

Dozens of early childhood observation systems are available and many of them have been used in a variety of settings and for a variety of purposes. (See, for example, Boyer, Simon & Karafin's Measures of Maturation: An Anthology of Early Childhood Observation Instruments, 1973, described in the Notes section, Appendix 1 of this booklet.) In the paragraphs below we describe five different general types of observation instruments. Also we will describe one example of each. Examples are given for illustrative purposes--not because they are necessarily recommended for general use. Indeed, observational approaches generally will have to be adapted for the particular use intended.

Continuous Records

It is impossible to observe and record everything that goes on in any classroom or social setting. Nevertheless, a continuous-record

approach to observation attempts to document relevant behaviors of a child, or events in a classroom, in a continuous, organized manner. Such behaviors can be recorded in narrative fashion or with some sort of checklist. Jane Stallings in her handbook Learning to Look (1977) describes how, as a teacher, she used a narrative continuous record to help understand and deal constructively with one troublesome youngster:

> Once, in desperation, when I could not understand the behavior of Billy, a particularly disturbing second-grade child, I hired a college student to come in and write a running account of everything he did for two days. From this, I received sixty hand-written pages of narrative.
>
> The information was most valuable. I learned that on the first day, Billy had gotten up and wandered about the room fifty-seven times. Since the school day was five hours long, this was about ten times an hour. He had fallen off his chair fourteen times. He had picked his nose seventeen times and rubbed his eyes twenty-three times. He had received thirteen smiles from me and twenty-seven reprimands -- mostly to stop falling off his chair and pay attention. He initiated conversations with other children forty-four times, but the interaction was only one or two sentences long. He spoke to everyone who passed his seat and tried to trip three people, succeeding twice. He was rejected fifteen times by other children who were involved in some activity and was physically pushed away from a group of three who were working on a mural. During recess, he put a blanket over his desk, took his reading workbook, and disappeared underneath. He stayed there for five minutes. The second day's observations were similar, and the picture that emerged was one of a hyperactive, highly distractible child.
>
> Supported by these specific descriptions, I requested conferences with his parents, his doctor, a reading specialist, and the school psychologist. The written account of his behavior enabled me to present factual information with a minimum of inference. As a result of these meetings, an educational program was planned that helped Billy progress in his learning. (p. 9).

Time Sampling

Continuous recording obviously can be an expensive and time-

consuming approach to observational assessment. An alternative is to use a time-sampling approach, under which observations are made at specified time intervals. The key ingredients of a time-sampling observation system are:

- The behavior or trait to be observed is defined in operational terms (specific actions or conditions).
- A time unit of observation (ranging from as little as one second to 15 minutes or more) is specified.
- A sampling strategy is specified (for example, observations might be made for the first 10-minute interval of each hour of the day).

A number of problems arise in applying such an observational strategy, of course, but since most of them are common to other observational techniques, let us postpone that discussion. Instead we simply illustrate this technique by describing an early application of a time-sampling approach used by Ruth Arrington (1932, also described in Wright, 1960, and Hutt & Hutt, 1970). Arrington's research concerned the behavior of young children. Her observational system was based on two checklists concerning activities which engaged children (use of materials, physical activity or no overt activity) and their social interactions (talking with others, non-social vocalizing, physical contact, laughing, or crying). These categories were defined to be mutually exclusive in terms of overt behavior of children. Individual children were observed during free play periods, using five-minute observation sessions during which children's activity engagement and social interactions were recorded every five seconds using special checklist forms (see Hutt & Hutt, 1970, for an example of Arrington's checklist forms). Checklist records were then analyzed to determine the frequency with which

different sorts of behavior occurred for individual children or different types of children. Arrington found, for example, that for nursery school children, non-social vocalizing was more frequent than social speech, and that children tended to converse primarily with members of their own sex.

Event Sampling

Like time sampling, event sampling can be more efficient than continuous recording. However, instead of observing and recording events in terms of a prespecified time sample, event sampling focuses on prespecified types of events or behavior. For example, such an approach might focus on question-asking behavior of children, or specific types of social interaction or their use of a play area.

Goodwin & Driscoll (1980) describe an event-sampling procedure employed in Kounin's (1970) study of kindergarten teachers' handling of classroom misbehavior during the first few days of school. In this study, the focus of observation was teachers' efforts to stop misbehavior, or what was called a desist. In addition to this primary event, observers also recorded information concerning the influence of the incident on neighboring children.

> When a teacher directed a desist at a misbehaving child, the observer recorded what the deviant child had been doing as well as activities of the audience (other children looking on), the nature of the desist and the deviant child's immediate reaction, and the behavior for the next two minutes of the nearest student witnessing the desist...observers waited until after the event to record particulars but did so immediately afterward to help assure fidelity of memory.

> Subsequent analysis and interpretation of the data on the events [showed that]...the ripple effect did, in fact, occur. Children witnessing a desist on the first day of kindergarten showed more overt reaction than on following days. On the first day, incidentally, they were more likely to behave themselves, to conform, or to show behavior disruption after viewing a desist. Deviancy-linked children showed more conformity, non-conformity, and a mixture of both after witnessing a desist than did deviancy-free children, and they were more likely to decrease deviancy and increase conformity if the desist was high in firmness. Clarity of desist influenced both categories of children in the direction of conformity and was, in general, more a determiner of the nature of the ripple effect than was firmness. Although rough desists upset many children, their overall effect on conformity and non-conformity was slight.
>
> (Goodwin & Driscoll, 1980, pp. 122-123)

Trait Rating

A fourth general type of observational technique is trait rating. With this approach, an observer does not directly describe behaviors or events, but instead, after observing a child or a classroom for a period of time, rates a general trait or characteristic of what was observed. A kindergarten teacher, for example, after watching and working with a child during the course of the school year, might rate a child in terms of the trait of readiness to begin a particular type of reading instruction.

In one observational study, which was part of the national evaluation of project Follow Through (FT), for example, observers were asked to rate several dimensions of FT first grade classrooms. Using a Physical Environment Information form which was developed as part of SRI International's observational study of FT, ooservers coded information on various aspects of the classroom setting: presence and use of specific equipment, intructional materials, games and toys; whether the classroom has movable or stationery tables and chairs, whether children's seating is assigned or self-selected, and whether children are assigned to or

select their own groups (Stallings & Kaskowitz, 1974, pp. 23-25). Subsequent analysis showed that the ratings could be used to discriminate reliably between classrooms affiliated with different FT model sponsors, and that some of these ratings were significantly related to children's later behavior in school and on tests.

Work Samples

A final type of observational technique is even less direct than the approaches described so far. Instead of observing children's behavior or classroom events directly and recording or rating them, this approach relies upon the collecting or recording of specimens of children's work; for example, drawings or other artwork and written materials. Again, we should point out that this form of assessment is by no means anything new. For decades teachers of young children have regularly sent children home with samples of their artwork and writing, as a means of helping parents appreciate what children have been learning. What is not so often recognized, however, is that such work samples also have potential value for systematic assessment.

Carini (1978) provides an example of this in what she calls documentary processes. She points out that the "accumulated work of a child in a medium such as writing, painting or blocks can be a focus of discussion" for teaching staff and parents. She describes how she employs such documentation, as follows:

> The first step in the documentation and portrayal of a child is to arrange the diverse forms of data--records, children's work, interviews, etc. --in chronological order. The entire record is re-read several times and pieces of the child's work are selected for description through a reflective conversation. For example, for a child (Misha) for whom the motif of houses is pervasive in stories and drawings, a number of reflections were carried out including "hidden", "domestic", and "wild". These reflections were followed by detailed descriptions of specific pieces of work.
>
> Immersion in the records and in the work allows themes or headings to emerge....
>
> The initial charting is followed by an unspecified number of rechartings according to the motifs, mediums and themes suggested by the initial exploration. Some of these headings are refinements of earlier headings, while others cut through the data from subtler angles than the more global characterization of the data provided by the initial headings....
>
> The last step in the study is the descriptive essay in which all of the data is integrated in order to portray the child. Stated concretely, the essay reflects the thematic patterns emergent from the records, and employs the particular data within the records to document those patterns.
>
> (pp.8-11)

Carini's systematic gathering and analysis of children's work samples together with other sorts of assessment information is quite unusual, but she explains that such methods of documenting and portraying children and their learning can prove extremely valuable.

> To portray the person to those primarily responsible for his or her education--teachers and parents--is to increase dramatically their capacity to make thoughtful choices in the interests of the child's education. At each point in the extended process described above, there is examination of setting, teaching practice, and the continuity of the child's experience and thought. It is also true that to see and know any one child fully is to know all children better. The uniqueness of the one calls up his or her shared perspectives with particular others, and embeds that perspective within the full range of human experience.
>
> (Carini, 1978, p.14)

POTENTIALS AND LIMITATIONS OF OBSERVATIONAL APPROACHES

As the examples cited above illustrate, there are several different sorts of approaches to observational assessment. As the examples suggested, these approaches need not rely exclusively on one type of sampling (for example trait or time sampling) but instead can combine sampling strategies. Also, any one approach is doubtless of limited use. Nevertheless, when applied in conjunction with other approaches, observational assessment has a tremendously broad range of uses. As illustrated in the examples we cited, systematic observational assessment may be of help to the teacher in planning instruction for individual children, to the researcher in charting the course of child development, to the evaluator in assessing the character, processes and outcomes of specific educational programs, and to the parent in understanding and promoting the learning of his or her child.

Nevertheless, observational approaches, like all forms of assessment, have weaknesses as well as strengths. First, we should point out that the same standards of technical quality pertain to observational techniques as to other forms of assessment. One must consider whether such observations are valid and reliable and provide a basis of comparison appropriate to the intended use. Validity of observations is important because research has shown that different observation systems that appear to measure the same sort of behavior can yield different results because of the way observation categories are defined or operationalized (Borich et al., 1977). Jane Stallings (1977) provides a specific example of the problem of obtaining reliable observations:

> The physical environment of the classroom--its size, shape, lighting, ventilation, and noise level--was considered important to the process of educating children. We tried to record this kind of information during our first two years of observation [of FT projects] but found it impossible to get observers to agree on what was "light enough" or "cool enough" or "quiet enough." Therefore, since we could not establish reliability among observers, we deleted these items from subsequent observations. (p.26)

Observational techniques for assessment have several other potential limitations which should be mentioned. For one thing, these approaches can be relatively expensive and time-consuming. Moreover, in order to produce valid and reliable measurement, special training of observers often is required. For example, before they are allowed to collect data using SRI International's Classroom Observation System for research purposes, observers are required to attend a seven-day training session and pass a criterion test (Stallings, 1977).

In addition to these practical limitations, observational approaches to early childhood assessment share a potential weakness common to all forms of assessment. The danger is simply that in focusing on observable behaviors or traits, or on available work samples, it is all to easy to let assessment become a goal in and of itself, concentrating on that which can be assessed easily, and to ignore broader issues in children's development and learning, forgetting the ultimate goal of how to promote that development and learning.

This chapter has provided only a very brief introduction to observational approaches to early childhood assessment. For references on sources of further information, see Appendix 1.

IV. USES OF EARLY CHILDHOOD ASSESSMENT INSTRUMENTS

It is difficult to evaluate evidence on the utility of a test or assessment procedure without considering the particular use to which the test or procedure is to be put. An assessment procedure may be good for some purposes but not at all for others. This point was made in several ways in the last two chapters. It is a simple notion, but one frequently overlooked in discussions of the technical quality of assessment procedures. Hence this section surveys alternative uses of educational assessment, and discusses them in light of the special issues of testing and assessment of young children.

First, however, we need to decide how to divide up the set of potential uses of assessment information for the sake of discussion. There are several ways one could do this. One reasonable way, suggested by a recent NIE report on testing, divides assessment use into four broad categories:

- To hold teachers, schools, and school systems accountable
- To make decisions concerning individual students
- To evaluate educational innovations and experimental projects
- To provide guidance to teachers in the classroom.

(White & Tyler, 1979, pp.7-8)

In the following pages we will discuss the special considerations bearing on use of early childhood tests and instruments for these four categories of use. Since use of assessment information for program evaluation is particularly salient with respect to Title I, we will discuss this type of use last, and in more detail than the others.

Also, since the distinction between norm-referenced and criterion-referenced assessment is relevant to types of use, let us spell out what is meant by these two terms. A norm-referenced test or assessment is designed to compare an individual's performance to that of others called a norming group or standardization sample. Criterion-referenced assessment is designed to compare an individual's performance not to that of other individuals, but to some other standard, such as a prespecified criterion score, or a domain of items or type of behavior. The distinction rests upon how assessment instruments are designed, not on how they are interpreted, since any test or assessment results can be interpreted in either norm- or criterion-referenced fashion. Thus one always should look beyond the labels of "criterion-referenced" and "norm-referenced" to investigate the content of an instrument and the exact manner in which it has been developed.

Assessment procedures for young children, for example, often are normed in terms of age rather than of grade level. Perhaps the most famous example of age-normed assessment of young children is Dr. Spock's The Common Sense Book of Baby and Child Care. The practice of age norming assessments of young children reflects two points noted earlier. First, before school entry the social and educational experiences of young children are diverse--hence there is no social or educational experience sufficiently common to most preschool children to provide a basis for norm-referenced test interpretations. Second, the age norms available for young children reflect the rapid development and change of children in their first five or six years of life. Gesell, Ilg and Ames's (1974) Infant and Child in the Culture of Today, for instance, provides behavior norms for the following ages: 4 weeks, 16 weeks, 28 weeks, 40 weeks, 1 year, 15 months, 18 months, 2 years, 2½ years, 3 years, 3½ years, 4 years, 4½ years, 5 years, 5½-6 years. The exact ages at which

certain behavior may be manifest will of course vary considerably with both individual characteristics and environmental influences, as Gesell et al. point out repeatedly. This variability is what makes the use of norms with young children so difficult. Research clearly suggests that not until around age nine (grade 3) has as much as 50 percent of the general achievement pattern at age 18 (grade 12) been developed (Bloom, 1964, p. 105). In other words, patterns of educational achievement are far more variable in the early childhood years (below grade 3) than in later years of schooling. From the assessment perspective, this suggests--as we said earlier--that measurement of young children is more difficult than that of older ones. Yet from an educational point of view, this finding has also been viewed as an opportunity. The great variability in young children's achievement and behavior has contributed to the theory that early childhood is a critical period for intervention--a time in which relatively minor alterations in environment can have immediate or long-term development consequences (White et al., 1973). But whatever its implications for educational practice, such variability makes norm-referenced interpretations of young children's performance particularly difficult. This in turn has implications for alternative uses of assessment information.

ADMINISTRATIVE AND PUBLIC ACCOUNTABILITY

As a recent NIE conference report on testing noted, educational assessment is used for a variety of accountability functions:

> Many principals, superintendents, and other education authorities use test scores, particularly scores on achievement tests, as a rough gauge of the adequacy of the performance of a teacher, a school, or a larger administrative unit. Parents, voters, and legislators also use such information in judging schools and school systems. The results of a test are taken to indicate the amount of learning accomplished by the average student in a classroom or larger unit.
>
> (White & Tyler, 1979, p.7)

As this account suggests, there are two major strands to the accountability functions of educational assessment--one for administrators and others directly and explicitly responsible for educational programs, and the other for parents and the public generally, who ultimately hold the authority for public education in the United States. The role of systematic educational assessment in both forms of accountability appears to be on the increase. In terms of administrative accountability, more and more educational programs require assessment of one sort or another. This is of course often tied to program evaluation functions, which will be treated later in this chapter.

The public accountability function of assessment, particularly standardized testing, has a longer history than formal program evaluation. Though testing has been explicitly tied to formal educational accountability schemes in recent years, test results have long served as a prime means by which the public judges the quality of schools. In some cities, newspapers have long published test results school by school. Real estate agents sometimes cite schools' test results to prospective buyers to entice them to buy homes in particular neighborhoods. Parents often are informed of their children's educational status in terms of test results.

In all such public accountability uses of educational assessment, there appears to be a strong tendency to rely on normative comparisons. One school's test results are compared to those of other schools. People want to know not just how many scholarships were awarded to seniors in high school A, but whether this was more or less than in other high schools in the area. Parents often want to know not just whether Johnny is doing okay in school, but how he is doing with respect to his peers. Desire for normative comparisons appears to be one important reason for the continuing

prominence of norm-referenced tests in educational assessment. One large city school superintendent, for example, was publicly asked why his schools continued to employ norm-referenced tests, despite the fact that they had developed an elaborate system of criterion-referenced assessment. He replied that the majority of taxpayers in his district, who do not have children in school, were not familiar with nor understood criterion-referenced results. "We show them norm-referenced results," he recounted, "to demonstrate the validity of what we are doing" (Haney, 1978, p.5).

This tendency in the accountability function of educational assessment probably also helps to explain the continuing use of grade-equivalent scores in American education. Experts in educational measurement have long warned against grade-equivalent scores because they are often misunderstood and misinterpreted (APA, AERA & NCME, 1974). Nevertheless, at least until recently, schools continued to rely heavily on grade-equivalent scores because they provided a familiar means of educational accounting. Grade-equivalent scores, despite serious problems of frequent misinterpretation, seem to remain popular simply because, as one observer recently put it, people <u>think</u> they understand what these scores mean, even if they do not.

These issues have implications for the use of early childhood assessment results for accountability functions. First, because of the limitations, or in many cases the nonexistence, of early childhood norms, it may be hard to report and interpret early childhood assessment results for public consumption. For children aged 3 to 4 or younger, age norms may provide a useful framework for interpreting assessment results. Yet by age 4 to 5, when children typically experience their first formal schooling, use of age norms becomes more hazardous. As we noted in Chapter 2, early educational

experience can sharply affect young children's educational performance. Unless this is taken into account, assessment results may inadvertently reflect the presence or absence of such experiences. Hence, normative comparisons, commonly made for accountability purposes at later grade levels, can be extremely difficult, if not altogether impossible, to carry out in a reasonable way at the early childhood level. One way to get around this problem is to report assessment results directly in terms of the assessment tasks employed--for example, instead of reporting normatively that children scored at the 70th percentile on a letter recognition test, to report in criterion-referenced fashion that 75% of them could recognize at least 20 letters of the alphabet.

A second and related issue in accountability uses of early childhood assessment has to do with the object of accountability. When high school students cannot read, or conversely, when they win numerous scholarships to college, this clearly reflects something about the schools they attend. Yet when young children, say in kindergarten, lack certain skills or are proficient in particular ways, it is often unclear to what extent this should be attributed to educational programs, to children's home and family background, to the particular characteristics of the children involved, or to other factors.

In short, the potential use of early childhood assessment for general accountability purposes--at least in ways traditionally used with standardized test results--seems to be somewhat less than that of assessment of older children. While there is not much good evidence on this point, several aspects of early childhood testing and assessment make this contrast plausible. Two points, discussed above, are: 1) the various problems of using norms with early childhood tests, and 2) the intertwined responsibilities of educational institutions and home and family for the early educational

development of young children. This suggests that alternative approaches to accountability may be useful at the early childhood level: approaches which seek to describe children's educational performance directly, rather than assessing it normatively or attributing causes for the performance.

MAKING DECISIONS CONCERNING INDIVIDUAL STUDENTS

Assessment results are also used to inform a range of decisions on individual students. At school entry they are used to help determine whether children are ready for reading instruction, or should be placed in special classes for the retarded or the gifted. Later in children's education, test results may be used to determine their eligibility for special programs such as Title I, and to assign them to different curriculum tracks in high school. Later still, in college or in the labor market, assessment results may affect admission or hiring and promotion decisions. Thus educational assessment plays a part in decisions about individual students throughout their educational and working careers.

Note that we are referring here only to major decisions concerning educational placement, promotion, and admissions--not to the shorter-term and less formal decisions, such as instructional guidance, which will be discussed separately in the next section. Nevertheless, even when we restrict attention to major educational decisions, the use of assessment results appears to be increasing. Within the last few years, for instance, numerous states have begun competency testing programs to control grade-to-grade promotion or to provide a basis for awarding high school diplomas.

These practices raise several issues. In selection decisions for college or jobs, the use of tests has traditionally been justified by demonstrations

that they have predictive validity--for example, that a college admissions test could predict student grades in college, or that job selection test results correlated with actual job performance. In the past, much has been written on issues of predictive validity, and particularly on bias in selection tests in terms of differential predictive validity.*

In the past few years, however, discussions on the use of assessment results for making analogous decisions about students at earlier levels in the educational system have taken a somewhat different direction. Rather than worrying simply about how well assessment results predict the performance of those selected for special opportunities, people involved in making decisions on selection and assignment of younger children have become more concerned with the consequences of selection, for those not selected as well as for those who are. In special education, for example, concern for both the negative and positive consequences of selecting and not selecting children for special programs has prompted enthusiasm for mainstreaming--that is, the integrating of children with special needs into regular classrooms, instead of segregating them in separate classes and possibly thereby stigmatizing them (Hobbs, 1975; Wolfensberger, 1972). Also, in the recent literature on competency testing, doubt has been raised about consequences of using such tests to promote students from grade to grade or to make them repeat a grade. In this light critics ask not just how well tests predict how children will do in the future, but how well they match what children have been taught in the past, and how much they help to improve what they learn in the future.

These views on use of assessment results for making decisions about individual students have special relevance at the early childhood level.

---

* See the Journal of Educational Measurement, 1976, Volume 13, for some good articles and references on this topic.

Readiness testing, for example, has a long tradition in early childhood education in America, but commonly used readiness tests actually cover very different sets of skills. Research suggests,too, that contrary to common opinion, separating "unready" children into transition classes, for example special kindergarten/first-grade classes, may not enhance their learning (Leinhart, 1980).

Before one can sensibly assess which readiness test to use, one must ask, readiness for what? The appropriateness of a given test to inform placement decisions will vary depending on the educational programs concerned. Also, the practical problems of assessment with young children described in Chapter 2 all caution against over-reliance on test results in making major decisions about young children. Because of these considerations, the following guidelines, widely accepted with respect to test use generally, are especially pertinent to the use of assessment results in making decisions about young children:

- A test user should consider more than one variable for assessment, and the assessment of any given variable by more than one method.
- A test user, in interpreting an obtained score, should consider the total context of testing before making any decisions (including the decision to accept the score).
- A test user should consider alternative interpretations of a given score.

(APA,AERA & NCME, 1974)

These guidelines serve to reemphasize the point noted earlier. Instead of relying simply on one form of assessment for making decisions about educational placement of young children, one should take into account alternative forms of assessment.

GUIDANCE TO TEACHERS IN THE CLASSROOM

A third class of uses of systematic assessment is to guide instruction--that is, to provide information and feedback to teachers as opposed to informing major administrative decisions. It is in this domain of use that early childhood assessment appears to be potentially most useful; at least this was suggested by a recent nationwide survey asking teachers how they used standardized achievement test results in their classrooms. More than 50 percent of responding kindergarten to grade 4 teachers replied that they used test results in only four of the ways the survey suggested:

Diagnosing strengths and weaknesses	77%
Measuring student growth	71%
Individual student evaluation	65%
Instructional planning	52%

(Beck & Stetz, 1979, Table 4)

Nevertheless, though tests appear to be relatively useful in guiding instruction, some observers have been highly critical of their usefulness for this purpose. The recent NIE report Testing, Teaching and Learning, for example, recounted the following:

> Several national educational groups have called for a moratorium on testing. It is argued that standardized tests have no positive direct usefulness in guiding instruction, and their indirect influence--implicitly laying down goals and standards--disrupts or blocks teaching. Despite inclusion in the published tests of various subtests to identify a student's strengths and weaknesses, critics say the categories are so broadly defined, the tests are given so infrequently, and the time from test administration to report of results to teachers is so long that tests do not help teachers in their work.
>
> (White & Tyler, 1979, pp.9-10)

Such criticism suggests several characteristics that may make assessment

results more useful in guiding instruction. First, they must be relevant to the goals of instruction--they must have what we called instructional validity in Chapter 2, Second, they must provide specific and accurate information on particular aspects of student learning. Third, they must provide feedback to teachers within a short time.

The first two characteristics--instructional validity and specificity of results--are two of the prime concerns behind the growth of interest in criterion-referenced testing within the last decade. In his recent book criticizing norm-referenced testing and advocating criterion-referenced testing, James Popham, for example, argued as follows:

> - Excessive generality in norm-referenced achievement tests leads to unrecognized mismatches between what is tested and what is taught.
> - Insufficient cues are available from norm-referenced test results to remedy ineffective instructional programs.
>
> (Popham, 1978, p. 84)

These concerns are obviously relevant to the use of early childhood assessment to guide instruction. If it is to be so used, assessment must 1) be matched to the goals of instruction, that is, have instructional validity; 2) provide specific information on individual children's strengths and weaknesses; and 3) allow rapid feedback of that information. Several of the special issues of early childhood assessment bear on these considerations. Readiness tests, for example, cover a range of skills that often are included within the goals of early childhood instruction; but as noted in Chapter 2, different readiness tests cover very different sets of such skills. In terms of match with goals of instruction, early childhood assessment is particularly weak in one area: social and emotional development, an important domain of early childhood instruction. As Walker noted

in her book Socio-emotional Measures for Preschool and Kindergarten Children:

> Very few [such] instruments have adequate standardization norms that are representative for a wide range of children of varying ethnic groups, intelligence levels and socio-economic backgrounds. Generally the ones that do exist are very poor and inadequate since they are based on extremely small or narrowly defined populations of children.
>
> (Walker, 1973, p. 37)

It is because of the weaknesses of paper-and-pencil measures of children's socio-emotional characteristics that Walker suggests the potential value of observational techniques of the sort described in Chapter 3.

In at least one respect, however, early childhood instruments may have more potential than later-grade tests for providing information useful in guiding instruction. As noted in Chapter 2, many early childhood assessment instruments are individually rather than group administered. When they are individually administered by the classroom teacher, she or he gains specific information immediately, even before scoring the test. If the information is keyed to particular goals of instruction, it can be immediately useful to the teacher in planning instruction. Thus, for the purpose of instructional guidance, early childhood assessment when keyed to instructional goals and administered individually appears to have more potential utility than group administered tests, the results of which may not be available to teachers until weeks after the said tests are given.

EVALUATION

A fourth class of assessment use is for evaluating educational programs and innovations. It is probably in this area that there has been the greatest increase in systematic educational assessment within the last two decades. As the NIE report Testing, Teaching and Learning put it:

> Government agencies, private foundations, and school systems sponsor experimental projects in American schools and seek to evaluate these projects through use of standardized achievement tests. A recent wave of experimental projects was the curriculum reform movement in science and mathematics, which began in the 1950's. Another, larger wave came in the 1960's when widespread efforts were made to improve the education of children from backgrounds of poverty and discrimination. Evaluators of experimental projects continue to wrestle with the task of matching tests to project objectives. In some cases, experimenters have found available tests unsuited to their projects and have developed new ones.
>
> (White & Tyler, 1979, p.8)

As this account suggests, the increased use of standardardized tests in program evaluation has not proven altogether satisfactory. Several observers, in fact, have directly criticized the widespread use of norm-referenced standardized tests in program evaluation (among others, Carver, 1974; Popham, 1978; Madaus et al., 1979). Their argument, in abbreviated form, goes roughly as follows. Norm-referenced tests were designed, historically, to serve selection purposes and hence to discriminate efficiently among individual test takers. As such they have been constructed to be insensitive to effects of instruction in local school systems, which may have different curricula. Now tests are increasingly being used to evaluate educational programs and to guide instruction. However, precisely because of the way they are constructed, norm-referenced tests tend to be insensitive to the instructional effects of particular educational programs. Hence new types of tests are required for the purposes of program evaluation.

More extreme critics of norm-referenced tests have extended this argument; they predict that the weaknesses of norm-referenced tests will usher in a new period of ecucational assessment-- "the criterion-referenced

measurement era" (Popham, 1978, p.2, emphasis in original). More moderate observers have suggested merely that curriculum-sensitive tests can play an important role in program evaluation, even though norm-referenced tests may continue to play a valuable role in comparisons of the educational outcomes of programs that emphasize different aspects of instruction (Madaus et al., 1979).

These and other criticisms have frequently been leveled against recent efforts to evaluate program impact. Five of the most common made with respect to early childhood programs are the following:

- There is often a real mismatch between the broad goals of early childhood educational programs and narrow test-based evaluations of them.
- There is often a great discrepancy between the long-range goals of early childhood programs (e.g., to prepare children to learn more in later schooling) and the short-term nature of most impact evaluations of them (e.g., end-of-program test scores).
- There has been a widespread failure to adequately describe the educational programs being evaluated, and to determine whether or not the program ostensibly being evaluated actually was implemented as intended.
- Most impact evaluations of early childhood programs yield few if any clearcut findings.
- Because of these problems among others, few impact evaluations provide information which is of much direct use in decision making or in improving programs.*

These criticisms obviously raise issues well beyond the mere use of tests in evaluating early childhood educational programs. Indeed, for that

---

* For more information on such criticisms with respect to past evaluations of early childhood educational programs, see Haney et al., 1978, pp.32-46.

reason, and because impact evaluation encompasses far more than simply testing and assessment, a range of issues in early childhood program evaluation is treated separately in other resource books in this series.

Nevertheless, several points should be made here with respect to using early childhood assessment instruments for program evaluation. First and foremost, the degree of match between early childhood programs and the test or tests used to evaluate them must be considered. While a case certainly can be made for testing aspects of children's development that are not encompassed in program goals, this should not be done inadvertently, for unintended mismatches between program goals and test content may affect evaluation results in misleading ways. This is especially true of early childhood assessment, where some common goals--for example, in the social and emotional domain--cannot be measured well with available tests, and in particular with paper-and-pencil tests. For this reason, observational techniques like those described in Chapter 3 may be especially valuable for early childhood program evaluation.

Second, since the main aim of educational evaluation, as opposed to educational research, is to inform decision making and to improve educational programs, one should closely consider the exact purpose an evaluation is to serve before selecting an assessment instrument. This applies not only to the content of instruments, but also to their form and to the way in which results can be derived from them. At the early childhood level, as we suggested in Chapter 2, the form of assessment (e.g., whether individually or group administered) can significantly affect results. In terms of how test results are reported for purposes of program evaluation,

one should be particularly careful with norm-referenced results. This caution is especially important at the early childhood level, because of the problems of norming tests with young children. The previous educational experience of children tested must closely match that represented in the norm group, or else norm-referenced results can be badly misleading.

In summary, if early childhood tests and instruments are used in evaluations of early childhood programs, one must give close attention not only to the special issues of early childhood assessment, but also to the particular goals of the programs to be evaluated.

USING EARLY CHILDHOOD ASSESSMENT FOR MULTIPLE PURPOSES

Assessment can serve many different functions. In this chapter, we have reviewed four different classes of such functions, dealing with:

- Accountability
- Making decisions about individual students
- Providing guidance for instruction
- Program evaluation.

As we pointed out at the start of the chapter, these distinctions are somewhat arbitrary. Program evaluation, for example, often serves accountability functions, and sometimes provides useful guidance regarding instruction, if not for individual students, at least regarding instructional practices at the classroom, school, or district levels.

Nevertheless, though they sometimes overlap, it is clear that different functions may require different forms of testing and assessment--or at least that different functions may pull assessment in different directions. For

some instructional purposes, for example criterion-referenced results, work samples or other observational approaches may be more useful than norm-referenced results, but for accounting to the public, norm-referenced results sometimes may be more useful. For some kinds of program evaluation, one might for technical reasons choose an assessment that yields results in the form of a standardized metric, whereas such a metric might be totally useless for public accountability functions. For purposes of evaluating social goals of early childhood programs or for research reasons, observational approaches may be especially useful even though they may prove cumbersome or too expensive for other purposes. Such contrasts suggest that different types of assessment should be used for different functions, or at least that, if one assessment is to serve different functions, it may have to be used in different forms, or the results reported in different ways.

These points will be summarized in Chapter 4. Here, let us briefly describe two issues that are relevant to systematic assessment for any function: the related issues of test bias and assessment with children who do not speak English as their first language.

Cultural Bias

Many people believe that standardized tests are biased against black and other minority children. A recent incident highlighting this concern was the finding by a federal court judge that standardized intelligence tests used in the state of California are racially and culturally biased and discriminate unfairly against black children. Ruling that intelligence tests have not been validated for the purpose of essentially permanent placement of children into educationally dead-end, isolated, and stigmatizing

classes for the so-called educable mentally retarded, the court enjoined the state of California from using IQ tests to place black children in such classes (Larry P. v. Riles, 1979, pp.3-4).

This ruling, though it applies legally only within the northern judicial district of California, nevertheless clearly highlights the widespread concern that standardized tests may be biased against minority children. There remains disagreement, however, over how to tell whether or not a particular test is biased. Different experts have proposed different definitions of test bias and different statistical methods for detecting fair use of tests (see Flaugher, 1978, and Petersen & Novick, 1976, for a review of these two issues respectively).

In light of the continuing debate over test bias, it is hard to propose specific remedies for this problem. Nevertheless, two general suggestions are appropriate. First, in selecting any test or assessment procedure for use with young children, one must consider whether its content and form are appropriate to the children's culture and background. Second, statistical analyses of test results may be irrelevant to issues of test bias if they ignore how assessment results actually are used. In other words, like validity, bias cannot be clearly determined in the abstract without taking into account how and with whom the assessment is to be used.

Language Considerations

A particular form of the general problem of cultural bias in assessment is the issue of assessment of children whose native tongue is not standard English. This problem has most often been discussed with respect to Spanish-speaking children, but is obviously relevant to any children who do not speak standard English as their native tongue.

There is growing awareness of the importance of bilingual education for such children and this awareness often extends to include concern for culturally sensitive assessment of children who do not speak English as their first language. Many people, for example, now recognize the importance of conducting assessments in the native language of the child if valid conclusions about his or her general educational development are to be drawn. There is not space here to treat issues of bilingual assessment in any detail (see Padilla, 1979, for a good recent survey of the literature on testing of Hispanic Americans). Nevertheless, a few general points can be mentioned. First, it is important to distinguish linguistic or cultural differences from other educational attributes, lest they be mistakenly interpreted as some sort of general learning deficit. Second, even when assessment is carried out in children's native tongues, the results of such assessments cannot be interpreted as being equivalent to those of English-language assessment; that is, merely translating a test into Spanish does not mean that its results with Spanish-speaking children are equivalent to results from the English version with English-speaking children. Third, issues of assessment with children who do not speak English as their native language must be viewed in light of the purposes of assessment. Using an English-language test with such children may be appropriate if the goal is to guide English-language instruction, but quite inappropriate if it is to measure children's general reading or math achievement. Fourth, although the problem of cultural bias in written language tests is widely recognized, it is often overlooked that assessment which relies on pictures may carry a problem of cultural dependency as great or even greater. Anastasi (1976), for example, argues:

> ...an item requiring that the names of the seasons be arranged in the proper sequence would be more appropriate in a cross-cultural test than would an item using pictures of the seasons. The seasons would not only look different in different countries for geographical reasons, but they would also probably be represented by means of conventionalized pictorial symbols which would be unfamiliar to persons from another culture.(p.347)

We have devoted special attention to the issues of test bias and assessment with children who do not speak English as their first language because these issues are particularly pertinent to early childhood testing and assessment. As we noted in Chapter 2, young children's performance on tests and other assessment tasks is easily affected by extraneous factors, including aspects of culture and language. Thus at the early childhood level, one needs to be especially attentive to potential cultural and language bias, regardless of the specific uses for which assessment is intended. How to minimize such problems is, of course, itself a problem. Nevertheless, in the next chapter we will offer some practical suggestions on how to deal with these issues in selecting and using early childhood tests and instruments.

V. SELECTING AND USING EARLY CHILDHOOD ASSESSMENT INSTRUMENTS

Given all of the potential problems in the testing and assessment of young children, how can one sensibly go about selecting and using an early childhood test or other assessment instrument? This is the question addressed in this section. We treat the question in three parts: screening potential instruments; trying out likely ones; and finally, using and interpreting results. The suggestions are often fairly general, for the simple reason that successfully selecting and using an early childhood test or observational instrument for any particular purpose will depend to a great extent on the specifics of that purpose, the conditions of assessment and the care with which results are considered and interpreted.

SCREENING POTENTIAL INSTRUMENTS

The primary points to consider in selecting any early childhood assessment device can be labeled as simply purpose and people. The first thing to consider is the exact purpose for which one intends to conduct an assessment. As noted in the last chapter, one assessment device may be good for some uses but altogether unsatisfactory for others. The second thing to be kept in mind is people: is the assessment procedure appropriate for use with the type of people--young children--with whom it is to be used? For example, group administered tests generally have limited validity for use with children below the age of six or seven.

With these points in mind, one should screen potentially useful instruments. Appendix 2 lists over one hundred early childhood test instruments and observation systems together with sources of additional information on each. This list is provided simply to illustrate types of instruments and give sources of further information. The fact that a particular instrument is listed should not be taken to mean that it is endorsed for any particular purpose, and the fact that

an instrument is not listed should not be taken to mean that it ought not be considered.

Screening of potentially useful instruments can be conducted efficiently in two steps. As an initial step, one needs only to review basic descriptive information for instruments that seem potentially useful. Examples of information for such initial screening are given in Appendix 3. As suggested in this appendix, initial screening of instruments can be accomplished simply by examining five characteristics of potentially useful instruments; namely, the type of instrument, the use intended for it by the publisher or developer, the population for which it is intended, its format, and its content.

Candidate instruments which seem potentially useful in terms of these characteristics can then be subjected to a more intensive review. Specifically one should screen potentially useful tests and instruments with respect to four categories of information:

- General information regarding the type and intended use of the instrument
- Theory, construction, and development of the instrument
- Practical requirements of the instrument
- Technical qualities of the instrument.

Table 1 outlines the kinds of specific information under these categories that ought to be considered in choosing assessment instruments. Appendix 4 provides some examples of detailed instrument reviews in format (again, however, the fact that particular instruments are reviewed in Appendix 4 should not be construed as an endorsement of them). Here let us simply explain the sort of questions which should be addressed with respect to each category of information, and why such questions are important.

As in making an initial review of early childhood assessment procedures which might be adopted, several sorts of general information need to be

Table 1: Outline of Information for Screening Assessment Instruments

Title:
Developer or Author(s):
Source or Publisher:
Copyright date or date of development:
Price:

I. General Descriptive Information

Type of instrument
Intended use
Intended population
Format
Content

II. Theory, Construction, and Development of Instrument

When and how instrument was developed
Manner in which items or assessment tasks were selected
Population or program for which instrument was developed

III. Practical Requirements

Materials required
Type of administration
Time and setting for administration
Directions
Sample questions
Scoring procedures
Language of administration
Training needed to administer

IV. Technical Information

Norms or other standards of comparison
Scales and scores
Validity
Reliability

V. Outside Reviews

Published reviews
Opinions of others who have used the instrument

VI. Comments

General
Theory, construction and development
Practical considerations
Technical qualities

VII. References

considered in making a detailed instrument review. What type of instrument or procedure is it? For what types of use and populations was it developed? What is the format and content of the instrument? In considering answers to these questions, one needs of course always to keep in mind one's own intended purposes for undertaking early childhood assessment.

With respect to theory, construction, and development of assessment instruments, one needs to ask whether each of these aspects of an instrument is reasonable and compatible with the use to which one wants to put it. If an instrument was constructed in terms of a specific psychological theory which seems irrelevant to the intended use, then one may want to reject it. If a test was constructed so as to discriminate between individual test takers regardless of their educational background, it may not be terribly useful in evaluating a particular educational program. Finally, if the intended use for an instrument corresponds to one of the uses which the instrument developer or publisher intended, then one can probably have more confidence that the instrument is a reasonable choice.

In terms of practical requirements, one should consider the accessory materials available with the instrument and the requirements for administering and scoring it. Tests which allow marking of answers directly in the test booklet, for example, are almost always more appropriate for use with early elementary school children than are those in which answers are marked on a separate answer sheet. Similarly, observation scales or individually administered tests in which an adult records children's responses generally are more appropriate for young children who have not mastered clerical test-taking skills. Also, tests or assessments which can be administered in short sessions of 15 to 20 minutes are generally preferable to those which require longer administration periods. The exceptions, of course, are instruments

that are individually administered and allow some flexibility of administration to help maintain children's attention, and observation instruments that do not intrude directly on the children's activities. Scoring requirements may also influence whether or not a test is useful for a certain purpose. Tests that can be hand-scored by the teacher may be more useful in providing information for instructional guidance than those which are machine-scored and returned to the teacher only after delays of a week or more. On the other hand, however, instruments which are not scored simply right/wrong, but which entail some judgment and interpretation in scoring, may require training for those doing the scoring.

In terms of technical quality, one should consider the available evidence on the validity and reliability of the instrument and the characteristics of norms provided with the test, if it is norm-referenced. If the test is to be used for program evaluation, for example, one must carefully review its content in light of the goals and objectives of the program. Although several people have suggested schemes for assessing the degree of match between test and program (e.g., Porter et al., 1978; Hambleton et al., 1978; Walker et al., forthcoming), such specific procedures will not be equally relevant for all assessment purposes. In most cases, however, a test will be appropriate the more its content covers program content, and the less it covers material irrelevant to the program.

If an instrument is to be used for selection purposes, different sorts of validity evidence will need to be considered. If the goal is to select for special services children who are likely to have difficulties in later

schooling, one needs to look for evidence that the instrument has predictive validity--that its results may be useful in predicting later school achievement.

Reliability evidence likewise should be examined in light of the assessment purposes one has in mind. If a test is to be used to help in making decisions about individuals, one needs to be far more concerned about reliability evidence than if it is to be used merely as an indicator of group or school progress.

If one is thinking of using norm-referenced tests, then test norms need to be considered in light of both the specific purpose of assessment and the type of children who are to be assessed. In the words of the 1974 Standards for Educational and Psychological Tests:

> In norm-referenced interpretations, a test user should interpret an obtained score with reference to sets of norms appropriate for the individual tested and for the intended use.
>
> (Standard J.5)

One mistake commonly made in this connection is to assume that, because the norming sample includes some individuals who are like the individuals or group with whom a test is to be used, the test norms are therefore appropriate. Even if a norming group contains a ten percent sample of minority children, for example, the norms are not therefore necessarily appropriate for use with minority children. Instead, one should examine the general characteristics of the overall norming sample. As the test Standards put it:

> A test user should examine differences between characteristics of a person tested and those of the population on whom the test was developed or norms developed. His responsibility includes deciding whether the differences are so great that the test should be used for that person.
>
> (Standard J.5.3)

TRYING THE TEST OUT

After one or more likely instruments have been identified, it is important to try them out with a small sample of the children with whom they are to be used, and then to discuss with those children how they interpreted test questions or assessment tasks and why they reacted to them in the way they did. This sort of practice--that is, pilot testing tests or other assessment instruments-- is important: it is all too easy for adults to forget that young children perceive the world, including tests and other assessment tasks, quite differently than do adults. A try-out will help to make clear whether test directions are too complicated for young children to follow, or whether certain questions are easily misinterpreted by or misleading to young children.

Trying instruments out with children with whom they are to be used is often neglected, but in our experience it can be immensely valuable. Even if this practice is followed with only one test that has already been selected, the findings can be very helpful in interpreting results. Recall, for instance, the example cited in Chapter 2 (p. 7) in which talking with children revealed that they often identified an elephant instead of a bird as the picture that goes best with the word "fly," because they saw the elephant as Dumbo, the flying elephant. Another example was revealed recently in a pilot test of a first grade screening instrument in a southern state. In one question children were asked to give their home address, and scoring procedures called for children to receive full credit if they responded with both their street and street number. What the pilot test revealed, however, was that some rural children came from homes which had no street number. Hence, scoring procedures had to be revised to take that fact into account.

For examples of how talking with children about how they perceive and interpret test questions can be a useful means of pilot testing instruments, see

Circourel et al., 1974; Haney and Scott, 1980; and Haney et al., 1981. Each of these sources is described in Appendix 1.

USING AND INTERPRETING TESTS

Given the various uses of tests and assessment devices, it is hard to offer specific advice on how early childhood tests and instruments should be used and results from them interpreted. Perhaps the most authoritative source of general advice on this topic is the 1974 version of Standards for Educational and Psychological Tests. Unlike earlier editions, the 1974 edition contains a special section on standards for the use of tests.

These standards are relevant to a wide range of uses of testing and assessment. The full document Standards for Educational and Psychological Tests (APA, AERA and NCME, 1974) treats these in some detail. Anyone not familiar with the standards may wish to read the full document. Here, let us simply elaborate on some of the standards especially relevant to early childhood assessment.

Regarding selection of a test or other method of assessment, consideration should be given to assessment of any given variable or attribute by more than one way. This is particularly important in assessing young children, since their performance and behavior may be highly variable, and since they often lack certain test-taking skills. For example, it often is helpful to use formal testing or assessment procedures in conjunction with teacher observation or checklists.

Regarding administration and scoring, one should follow standard procedures relevant to the instrument employed along with procedures that enable each child to do his or her best. Again, this is especially important at the early childhood level. Since formal assessment may be threatening or at least

unfamiliar to young children, it is vital for the test administrator or the observer to establish rapport with children and to make sure that they feel comfortable in the assessment situation.

Regarding interpretation, one point in particular is relevant to the early childhood level. Assessment results should be interpreted as an estimate of performance under a given set of circumstances; they should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. Violation of this principle has probably led to more misuse of standardized testing with young children than any other. Children's performance may be influenced by behavior problems, visual or hearing defects, language problems, and ethnic or cultural factors. Thus, it is vital to consider the total context of testing or assessment in interpreting results. In general, one should avoid use of descriptive labels that might be misinterpreted. As the Standards points out:

> The use of a summary label connotes value judgments; unfortunately most are words used in everyday language and therefore subject to inaccurate interpretation. A test maker may know precisely what he means when he uses the term "retarded," but he has no influence over the interpretation of the same word by a judge, teacher, parent or child.
>
> (Standard J.2.3)

To help avoid problems of misinterpretation, such terms as grade-equivalent, IQ, or IQ-equivalent should be used with utmost caution, if at all. Both IQ scores and grade-equivalents involve severe technical problems. Serious misinterpretations often occur, for example, when grade levels are extrapolated beyond the range for which the test is designed. Moreover, many test users fail to recognize the wide margin of error implicit in IQ or grade-equivalent scores. Indeed, because of widespread misinterpretation and misuse of such scores, many experts recommend that neither IQ nor grade-equivalent scores be

APPENDIXES

AND

REFERENCES

Appendix 1

NOTES ON SOURCES OF FURTHER INFORMATION

Since this booklet has provided only a brief introduction to issues in early childhood assessment, this section provides notes on relevant sources of further information.

GENERAL SOURCES

One helpful source of a wide range of information on early childhood assessment is Goodwin and Driscoll's Handbook for Measurement and Evaluation in Early Childhood Education (1980). This volume provides: a review of basic measurement concepts; a discussion of validity, reliability, and usability of measures; a review of observational measurement in early childhood; and separate chapters on (1) intelligence and school-related tests; (2) developmental and handicapped screening surveys; language, bilingual, and creativity tests; (3) affective measures, and (4) psychomotor measures. In addition this handbook provides helpful reviews of (1) conceptual frameworks for evaluation; (2) several recent large-scale evaluations, and (3) relevant information from other fields such as sociology and anthropology.

Anastasi's Psychological Testing (4th edition, 1976) and Cronbach's Essentials of Psychological Testing (1970) are both excellent general texts on educational and psychological testing. Anastasi's book includes two brief sections devoted to early childhood testing and assessment, one on infant and preschool testing (p. 266) and another on intelligence in early childhood (p. 332).

Johnson's Preschool Test Descriptions (1979) describes 170 preschool tests in terms of identifying information, administration, examinee appropriateness, interpretation, technical aspects, and additional comments.

In terms of purpose, each instrument is described as emphasizing screening, diagnosis, or achievement.

Hoepfner, Stern and Nummedal's CSE-ECRC Preschool/Kindergarten Test Evaluations (1971) is another potentially useful source of information concerning early childhood instruments. This volume lists several hundred early childhood instruments (including both full test instruments and subtests). The instruments are organized into four broad areas concerning the affective domain, the intellectual domain, the psychomotor domain, and subject area achievement. Each test or subtest is rated via a point and letter rating system in terms of measurement validity, examinee appropriateness, administrative usability, and normed technical excellence. While the broad patterns of these ratings provide some useful information, for example showing that most instruments are relatively weak in terms of providing validity and other technical evidence, considerable caution should be exercised in interpreting specific ratings. For example, while Hoepfner and his colleagues apply a simple rating system to all tests reviewed, different possible applications call for different weight to be given to the various attributes of an instrument.* Hoepfner et al.'s CSE Elementary School Test Evaluations (1976) covers instruments appropriate for grades 1-6, but the caution suggested with respect to the 1971 volume is relevant to this volume also.

Johnson's Tests and Measurements in Child Development Handbook II (1976) describes nearly 900 unpublished measures of child behavior. Measures are classified into the following categories: (1) cognition, (2) personality and emotional characteristics, (3) perceptions of environments, (4) self-concept, (5) qualities of care given and home environment,

* See Haney et al., 1978, pp. 110-111 for a discussion of some of the drawbacks in the CSE approach to rating test quality.

(6) motor skills and sensory perceptions, (7) physical attributes, (8) attitude and interests, (9) social behavior, and (10) vocational. Listings for each measure include identifying information, description of the measure, reliability and validity information, and bibliography. An earlier edition of this book, organized along similar lines, was Johnson and Bommarito's Tests and Measurements in Child Development (1971) which listed around 300 unpublished instruments.

D. Walker's Socioemotional Measures for Preschool and Kindergarten Children (1975) describes 143 instruments designed to measure social and emotional measures of young children. Each is described in terms of identifying information, general description, norms, validity and reliability information.

Buros' Mental Measurements Yearbooks (MMYs) are clearly the single best general source of information on specific tests and instruments. Eight MMYs have been published since 1938, but the Sixth MMY (1965), the Seventh MMY (1972) and the Eighth MMY (1978) are the only volumes with information relevant to most currently used tests and instruments. Buros' Tests in Print I (1961) and II (1974) provide comprehensive indexes to previously published Yearbooks. Tests in Print II also includes a reprint of the APA, AERA, & NCME Standards for Educational and Psychological Tests (APA, AERA, & NCME, 1974). Buros' Yearbooks deal with a wide range of tests besides early childhood instruments, but the Seventh MMY, for example, describes more than 500 instruments appropriate for children in the prekindergarten to grade one age range. The Buros volumes are especially helpful in comparison to others because they provide critical reviews of most of the tests listed.

Other recommended sources on general issues in early childhood assessment are Bradley and Caldwell (1974) on issues of testing young children, Cazden (1971) and Kamii (1971) both dealing mainly with assessment and evaluation at the preschool level, Raizen and Bobrow (1974) concerning evaluation of social competence development in Head Start and White et al. (1973) concerning a wide range of federal programs for young children and research and evaluation of these programs.

Three sources recommended as examples of what can be learned by pilot-testing instruments on a small-scale basis prior to full-scale use are Cicourel et al, 1974; Haney and Scott 1980; and Haney et al. 1981. Cicourel et al. (1974), particularly Chapter 5, describes an analysis of how first grade students arrived at answers to a reciting test, on the basis of interviews with children after they had taken the test under standard conditions. Haney and Scott (1980) describe a similar analysis of how second- and third-grade children perceived and reasoned about reading, science and social studies test questions from four of the most commonly used standardized achievement test series. For a more specific example of how a readiness instrument was pilot-tested with kindergarten children and revised on the basis of pilot-study findings, see Haney et al., (1981), pp. 48-50.

OBSERVATIONAL APPROACHES TO EARLY CHILDHOOD ASSESSMENT

A variety of sources of information regarding observational approaches to early childhood assessment are available.

Almy and Genishi's Ways of Studying Children (1979) subtitled "An Observational Manual for Early Childhood Teachers," provides a good discussion of alternative ways of observing children. Specifically discussed are studying the way children think, asking children about themselves, studying children in groups, studying the ways children express themselves, and studying

the child through others. Though the book is aimed primarily at teachers, it also would be of value to anyone interested in observational approaches to child study.

Boehm and Weinberg's The Classroom Observer (1977) provides a good introduction to systematic classroom observation. This book aims at helping readers "derive valid and reliable information about children in their natural habitat through the correct and relevant use of observational strategies" (p. xi). After an introduction concerning the selective nature of observation, the book discusses (1) defining the problem and describing the setting, (2) labeling and categorizing behavior, (3) sampling and recording behavior, (4) the teacher as observer, (5) the relationship between media and observation and (6) applying observation skills to education.

Borich and Madden's Evaluating Classroom Instruction: A Sourcebook of Instruments (1977) reviews almost 170 instruments relevant to evaluation of classroom instruction. These include rating scales, checklists, observational coding systems, and self-report questionnaires. A variety of types of instruments are reviewed because the authors seek to "encourage multivariate methods of research" (p. 6). Instruments reviewed are organized according to who (teacher, pupil, or observer) provides information about whom (the teacher, the pupil, or the classroom). Each instrument is described in terms of general information and description, illustration of sample items and response functions, psychometric characteristics, norms, administration and scoring, comments and references. Though the range of instruments described are not limited to the early childhood levels, several sections of the book (especially IIA About the Pupil from the Teacher, IIC About the Pupil from an Observer, IIIA About the Classroom from the

Teacher and IIIC About the Classroom from an Observer) describe instruments relevant to early childhood assessment.

Boyer, Simon, and Karatin's Measures of Maturation: An Anthology of Early Childhood Observation Instruments (1973) describes more than 70 observation systems for use in observing and recording behaviors of infants and young children. Each is described in terms of rationale and purpose, dimensions of the system, instructions for use, and references and related research.

Carini's monograph The Art of Seeing and the Visibility of the Person (1979) describes "a metaphysics of observing and presents a method for gathering and organizing empirical observation in order to disclose meaning" (p. 7). Rather than focusing on particular observational techniques, this monograph aims at describing the art of observation and reflection on children through time so as to derive protrayals that "disclose the continuity and transformation in [their] thinking as these are revealed in their projects and activities, in such . . .mediums of expression as drawing, building, and writing."

Goodwin and Driscoll's Handbook (1980), described in general above, provides a useful introduction to observational approaches to early childhood assessment in Chapter Four. The rationale behind this chapter is that "Carefully conceptualized and applied observational procedures can complement other measures available for use in various settings" (p. 111). This chapter outlines the importance of observational measurement in early childhood assessment, describes formal and informal approaches to observational measurement, recounts the general advantages and limitations of observational measurement, and illustrates three hypothetical applications of observational measurement.

Stallings' Learning to Look, A Handbook on Classroom Observation and Teaching Models (1977) provides another useful introduction to observational assessment in general and to one observational instrument in particular, the SRI classroom observation system. This system was developed in the course of the national evaluation of Project Follow Through, and consists of three instruments: the physical environment interaction form, the classroom checklist, and the five-minute observation form. Stallings' book also describes five different models of early elementary education of the sort included in the Follow Through Program (the exploratory, group process, developmental, cognitive, programmed, and fundamental school models) and briefly reviews evidence from the Follow Through evaluation on how children grow and develop in each of those models.

On more technical issues regarding observational measurement generally, see Garner (1960), Guilford et al. (1962), Wright (1967), Medley and Mitzel (1963), Hutt and Hutt (1970), and Borich et al. (1977).

SPECIAL ISSUES

The reader may also wish to pursue some of the special topics mentioned in this booklet through other readings.

On criterion-referenced measurement, Popham (1978) presents a good introduction. Hambleton and Eignor (1978) describe a set of guidelines for possible use in evaluating criterion-referenced tests and test manuals. Berk (1980) and Hambleton et al. (1978) provide useful reviews of a variety of technical issues in criterion-referenced measurement.

Regarding the use of systematic measurement with special groups--ethnic minority children, those who do not speak English as a first language or otherwise special individuals--several sources are helpful. Miller (1974), Oakland (1977), and Hilliard (1979) provide useful reviews of issues in

the assessment of black and minority children generally. Padilla (1979) provides a similar review with respect to Hispanic Americans. Flaugher (1978) provides a useful review of the many definitions of test bias, and Petersen and Novick (1976) give a good review of alternative conceptions of fairness in selection testing. Hobbs (1975) presents a broader discussion of the use of assessment results in classifying and labelling of children.

OTHER SOURCES OF INFORMATION

All of the sources mentioned above are of somewhat limited value in that they are printed material, and as such may become outdated with the passage of time. Hence, let us also recommend several institutional sources which may be useful in that they provide information on a variety of topics on an ongoing basis.

ERIC. The Educational Resources Information Center (ERIC) network is one of the most valuable of such sources of ongoing information. The ERIC system encompasses a computerized information retrieval system covering a wide variety of educational materials, both published and unpublished. A description of the ERIC system is available in NIE's publication ERIC: A Prolife, and suggestions on how to use the ERIC system are provided in Brown, Sitts, and Yarborough (1975) and Simmons (1975). The ERIC system is based on 16 ERIC clearinghouses which collect, evaluate, and distribute information concerning particular topical areas. Three ERIC clearinghouses relevant to early childhood assessment, with notes on the scope of areas they cover, are:

ERIC Clearinghouse on the Disadvantaged
Columbia University, Teachers College
Box 40
525 W. 120th Street
New York, New York 10027
Telephone: (212) 678-3780

Effects of disadvantaged experiences and environments, from birth onward; academic, intellectual, and social performance of disadvantaged children and youth from grade 3 through college entrance; programs and practices provide learning experiences designed to compensate for special problems of disadvantaged; issues, programs, and practices related (1) to economic and ethnic discrimination, segregation, desegregation, and integration in education; and (2) to redressing the curriculum imbalance in the treatment of ethnic minority groups.

ERIC Clearinghouse on Early Childhood Education
University of Illinois
College of Education
805 W. Pennsylvania Avenue
Urbana, Illinois 61801
Telephone: (217) 333-1386

Prenatal factors, parental behavior; the physical, psychological, social, educational, and cultural development of children from birth through the primary grades; educational theory, research, and practice related to the development of young children.

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service
Princeton, New Jersey 08540
Telephone: (609) 921-9000 ext. 2182

Tests and other measurement devices; evaluation procedures and techniques; application of tests, measurement, or evaluation in educational projects of programs.

More general information on the ERIC system and its other clearinghouses is available from:

Educational Resources Information Center
(Central ERIC)
National Institute of Education
Washington, D.C. 20208
Telephone: (202) 254-5040

ETC Head Start Test Collection. The Educational Testing Service also administers the Head Start Test Collection which was established to provide information about assessment instruments concerning children from birth to nine years of age. Qualified persons working in the area of early childhood education may have access to the collection in person or via mail or

phone inquiries. The collection also publishes a series of bibliographies on special early childhood assessment topics, which include:

- Self Concept Measures: An Annotated Bibliography (ED 051 305)
- Language Development Test: An Annotated Bibliography (ED 056 082)
- School Readiness Measures: An Annotated Bibliography (ED 056 083)
- Tests for Spanish-Speaking Children: An Annotated Bibliography (ED 056 084)
- Measures of Social Skills: An Annotated Bibliography (ED 056 085)
- Assessing the Attitudes of Young Children Toward School (A State-of-the-Art Paper) (ED 056 086)
- Measure of Infant Development: An Annotated Bibliography (ED 058 326)

For copies of these bibliographies or further information on the Head Start Test Collection, write to:

Head Start Test Collection
Educational Testing Service
Princeton, N.J. 08540

Title I Technical Assistance Centers (TACs). The TACs serving the ten regional areas of the United States are also sources of information on educational assessment, particularly with respect to Title I evaluation.

Region I: Connecticut, Maine, Massachusett, New Hampshire, Rhode Island, and Vermont

-RMC Research Corporation
400 Lafayette Road
Hampton, N.H. 03842
Telephone: (603) 926-8888
436-5385

Region II: New York, New Jersev, Puerto Rico, and the Virgin Islands

-Educational Testing Service
Princeton, N.J. 08540
Telephone: (609) 734-5117

Region III: Delaware, Maryland, Pennsylvania, Virginia, West Virginia, and the District of Columbia

-NTS Research Corporation
2634 Chapel Hill Blvd.
Durham, N.C. 27707
Telephone: (919) 493-3451
(800) 334-0077

Region IV: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee

-Educational Testing Service
Southern Regional Office
250 Piedmont Avenue
Suite 2020
Atlanta, Georgia 30326
Telephone: (404) 524-4501

Region V: Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin

-Educational Testing Service
1 American Plaza
Evanston, Illinois 60201
Telephone: (312) 869-7700

Region VI: Arkansas, Louisiana, New Mexico, Oklahoma, and Texas

-Powell Associates
3724 Jefferson
Suite 205
Austin, Texas 78731
Telephone: (512) 453-7288
(800) 531-5239

Region VII: Iowa, Kansas, Missouri, and Nebraska

-American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302
Telephone: (415) 494-0224

Regions VIII, IX and X: Colorado, Montana, North Dakoka, South Dakota, Utah, and Wyoming (Region VIII); Arizona, California, Hawaii, Nevada, Guam, Trust Territory of the Pacific Islands, and American Samoa (Region IX); and Alaska, Idaho, Oregon, and Washington (Region X).

-Northwest Regional Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204
Telephone: (503) 248-6853

Appendix 2

LISTING OF SELECTED EARLY CHILDHOOD INSTRUMENTS
AND SOURCES OF REVIEW INFORMATION ON EACH

This appendix lists over 100 early childhood assessment instruments and sources of review information on each one. For each instrument listed, the following information is given:

Title
Type
Publisher
Copyright date(s)
Grade or age span for which intended
Sources of review information on the instrument.

The titles listed are ones which have been indicated to have been used in ECT-I programs in the past, for screening, needs assessment or program evaluation. The listing of particular titles does not constitute agency endorsement of those instruments, nor does it mean that others should not be considered.

The types used to describe instruments are drawn mainly from the series of test review volumes written by Oscar Buros, the most widely known source of review information on tests. However, it should be noted that other authorities often use other typologies to describe types of tests. Some of the instruments listed by Buros as measuring personality characteristics, for example, often are described by others as measuring effective characteristics.

Since the publishers frequently update information on their instruments or produce altogether revised versions, at the back of this appendix we have listed the addresses of publishers who have issued tests intended to be useful for assessment at the early childhood level. If one is seriously

considering use of a particular instrument, it is advisable to write to the publisher to obtain up-to-date information.

The sources of review information on the instruments listed refer to the following publications.

T2: Buros, O. (Ed.). Tests in print II. Highland Park, N.J.: Gryphon Press, 1974.

7MMY: Buros, O. (Ed.). Seventh mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1972.

8MMY: Buros, O. (Ed.). Eighth mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1978.

CSE-ECRC: Hoepfner, R., Stern, C., & Nummedol, S. (Eds.). CSE-ECRC preschool/kindergarten test evaluation. Los Angeles, CA.: Center for the Study of Evaluation and Early Childhood Research Center, 1977.

CSE: Hoepfner, R. et al. CSE Elementary school test evaluations. Los Angeles, CA: Center for the Study of Evaluation, 1976.

J+B: Johnson, O., & Bomoncrito, J. Tests and measurements in child development. San Francisco, CA: Jossey-Bass, 1971.

OJ: Johnson, O. Tests and measurements in child development: Handbook II. San Francisco, CA: Jossey-Bass, 1976.

HJ: Johnson, H. Preschool test descriptions. Springfield, Illinois: Thomas, 1979.

W: Walker, D. Socioemotional measures for preschool and kindergarten children. San Francisco, CA: Jossey-Bass, 1973.

Numbers given for T2, 7MMY and 8MMY refer to test entry numbers. Those for J+B, OJ, HJ, and W are page references. No page references are provided for CSE and CSE-ECRE because in these volumes, information on particular instruments typically is spread across a fair number of pages.

EARLY CHILDHOOD ASSESSMENT INSTRUMENTS

ABC Inventory to Determine Kindergarten and School Readiness
Readiness
Research Concepts. [Educational Studies & Development]
1965
Entrants to kindergarten or grade 1
T2: 1691
7MMY: 739
CSE-ECRC: K6
J+B: 27-28
HJ: 25-26

American School Achievement Tests
Achievement
Bobbs Merrill Co., Inc.
1941-75
Grades 1, 2-3, 4-6, 7-9
8MMY: 4
CSE:
HJ: 31-32

Analysis of Readiness Skills: Reading and Mathematics
Reading Readiness
Houghton Mifflin Co.
1969-72
Kindergarten - 1
8MMY: 796
CSE:

Animal Crackers: A Test of Motivation to Achieve
Personality
CTB/McGraw-Hill
1973-75
Preschool - grade 1
8MMY: 497
CSE:

Basic School Skills Inventory
Miscellaneous: Learning Disabilities
Follett Publishing Co.
1975
Ages 4-6
8MMY: 424
CSE:
HJ: 45-46

Bayley Scales of Infant Development
Intelligence - Individual
Psychological Corporation
1969
Ages 2-30 months
8MMY: 206
7MMY: 402
HJ: 47-48

Bender-Gestalt Test
Personality
American Orthopsychiatric Association, Inc.
1938-48
Ages 4 and over
8MMY: 506
7MMY: 161
HJ: 51-52

Bilingual Syntax Measure
Foreign Language - Spanish
Psychological Corporation
1973-76
Bilingual children, kindergarten - grade 2
8MMY: 156
CSE:

Boehm Test of Basic Concepts
Intelligence - Group
Psychological Corporation
1967-71
Kindergarten - grade 2
8MMY: 178
7MMY: 335
CSE-ECRC:
CSE: 36
HJ: 55-56

Botel Reading Inventory
Reading - Miscellaneous
Follett Educational Corporation
1961-70
Grades 1-4, 1-6, 1-12
T2: 1658
7MMY: 727
CSE:

California Achievement Tests
Achievement
CTB/McGraw Hill
1934-74
Grades 1.5-2, 2-4, 4-6, 6-9, 9-12
8MMY: 10
7MMY: 5
CSE:

California Preschool Social Competency Scale
Personality
Consulting Psychologists Press, Inc.
1969
Ages 2.5 - 5.5
8MMY: 513
7MMY: 48
W: 261-262
CSE-ECRC:
HJ: 67-68

California Short Form Test of Mental Maturity
Intelligence - Group
CTB/McGraw-Hill
1938-65
Kindergarten - 1.4, 1.5-3.4, 3-4, 4-6, 6-7, 7-8, 9-12, 12-16, adults
8MMY: 179
7MMY: 337
CSE-ECRC:
CSE: 32

Children's Embedded Figures Test
Personality
Consulting Psychologists Press, Inc.
1963-71
Ages 5-12
8MMY: 519
7MMY: 53
CSE:

Children's Self-Social Constructs Test
Interests or Preferences
Virginia Research Associates
1967
Preschool
W: 141-142
HJ: 81-82

Circus
Readiness, language, and motor development
Addison-Wesley Publishing Company, Inc.
1979
Preprimary to K.5 (Circus A), K.5 - 1.5 (Circus B), Pimrary to 1.5-2.5
HJ: 85-86

Cognitive Abilities Test
Intelligence - Group
Houghton-Mifflin
1954-74
Kindergarten - 1, 2-3, 3-12
8MMY: 181
7MMY: 343
CSE-ECRC: K17
CSE:

Cognitive Skills Assessment Battery
Reading Readiness
Teachers College Press
1974
Prekindergarten
8MMY: 797

Columbia Mental Maturity Scale
Intelligence - Individual
Psychological Corporation
1954-72
Ages 3-6 to 9-11
8MMY: 210
CSE-ECRC
CSE:
HJ: 87-88

Comprehensive Identification Process
Miscellaneous - Learning Disabilities
Scholastic Testing Service, Inc.
1975
Ages 2.5 - 5.5
8MMY: 425
HJ: 91-92

Comprehensive Test of Basic Skills
Achievement
CTB/McGraw-Hill
1968-76
Kindergarten - 1.2, Kindergarten.6 - 1.9, 1.6 - 2.9, 2.5 - 4.9, 4.5 - 6.9, 6.5 - 8.9, 8.5 - 12.9
8MMY: 12
7MMY: 9
CSE:

Cooperative Preschool Inventory
Intelligence - Individual
Cooperative Tests & Services
1965-70
Disadvantaged children ages 3-6
T2: 490
7MMY: 404
HJ: 95

Cooperative Primary Tests
Achievement
ETS; Addison-Wesley Publishing Company, Inc.
1965-67
Grades 1.5 - 2.5, 2.5 - 3
8MMY: 13
7MMY: 10
CSE:

Denver Developmental Screening Test
Intelligence - Individual
Ladoca Project & Publishing Foundation
1968-70
Ages 2 weeks - 6 years
T2: 492
7MMY: 405
J+B: 32-33
HJ: 99-100

Detroit Tests of Learning Aptitude
Intelligence - Individual
Bobbs Merrill Co., Inc.
1935-75
Ages 3 and over
8MMY: 213
7MMY: 406
CSE-ECRC
CSE:

Developmental Test of Visual Perception
Vision
Consulting Psychologists
1966
Ages 3 - 8
8MMY: 882
HJ: 115-116

Developmental Tests of Visual-Motor Integration
Sensory-Motor
Follett
1967
Ages 2-8, 2-15
8MMY: 870
7MMY: 867
CSE-ECRC:
CSE:
HJ: 113-114

Diagnostic Reading Scale
Reading - Diagnosis
CTB/McGraw Hill
1963-75
Grades 1-6 and poor readers in grades 7-12
8MMY: 753
7MMY: 717
CSE

Draw-A-Person
Character - Projective
Western Psychological Services
1963
Ages 5 and over
T2: 1455
7MMY: 165

Durrell Analysis of Reading Difficulty
Reading - Diagnostic
Harcourt Brace Jovanovich, Inc.
1933-53
Grades 1-6
T2: 1628
CSE:

Durrell Listening-Reading Series
Reading - Miscellaneous
Harcourt Brace Jovanovich, Inc.
1969-70
Grades 1-2, 3-6, 7-9
T2: 1660
7MMY: 728
CSE:

Durrell-Sullivan Reading Capacity and Achievement Test
Reading - Miscellaneous
Harcourt Brace Jovanovich, Inc.
1937-45
Grades 2-5 - 4.5, 3-6
T2: 1661

Gates-MacGinitie Reading Tests
Reading
Houghton Mifflin Co.
1926-72
Grades 1, 2, 3, 2.5-3, 4-6, 7-9
8MMY: 726A
7MMY: 689
CSE-ECRC:
CSE:

Goodenough-Harris Drawing Test
Intelligence - Group
Psychological Corporation
1926-63
Ages 3-15
8MMY: 187
7MMY: 352
CSE-ECRC:
CSE:
HJ:

Gray Oral Reading Test
Reading - Oral
Bobbs-Merrill Co., Inc.
1963-67
Grades 1-16 and adults
2: 1681
CSE:

Illinois Test of Psycholinguistic Abilities
Miscellaneous - Learning Disabilities
University of Illinois Press
1961-68
Ages 2-10
8MMY: 431
7MMY: 442
CSE-ECRC:
CSE:
HJ: 138-139

Individualized Criterion-Referenced Testing
Reading - Diagnosis
Educational Development Corporation
1973-76
Kindergarten, 1, 2, 3, 4, 5, 6, 7, 8
8MMY: 764

Individualized Criterion-Referenced Testing
Mathematics
Educational Development Corporation
1973-77
Grades 1, 2, 3, 4, 5, 6, 7, 8
8MMY: 275

Iowa Tests of Basic Skills
Achievement
Hoghton Mifflin Co.
1955-73
Grades 1.7-2.5, 2.6-3.5, 3-9
8MMY: 19
7MMY: 481

Key Math Diagnostic Arithmetic Test
Mathematics - Arithmetic
American Guidance Service
1971-76
Preschool - Grade 6
8MMY: 305
CSE:
HJ: 146-147

Kindergarten Auditory Screening Test
Follett Publishing Co.
1971
Kindergarten - Grade 1
8MMY: 940
CSE:
HJ: 148-149

Leu-Clark Reading Readiness Test
Reading - Readiness
CTB/McGraw-Hill
1931-62
Kindergarten - Grade 1
T2: 1563
7MMY: 752
CSE-ECRC
CSE:

McCarthy Scales of Children's Abilities
Intelligence - Individual
Psychological Corporation
1970-72
Ages 2.5 - 8.5
8MMY: 219
CSE:
HJ: 172-173

Meeting Street School Screening Test
Miscellaneous - Learning Disabilities
Crippled Children & Adults of Rhode Island, Inc.
1969
Kindergarten - Grade 1
8MMY: 435
7MMY: 756
CSE:
HJ: 174-175

Metropolitan Achievement Test
Achievement
Psychological Corporation
1931-73
Kindergarten.7-1.4, 1.5-2.4, 2.5-3.4, 3.5-4.9, 5.0-6.9, 7.0-9.5
8MMY: 22
7MMY: 14
CSE:

Monroe Reading Aptitude Tests
Reading readiness
Houghton Mifflin
1935-63
Kindergarten - Grade 1
T2: 1724
CSE-ECRC:

Murphy-Durrell Reading Readiness Analysis
Reading Readiness
Psychological Corporation
1947-65
First grade entrants
8MMY: 803
7MMY: 758
CSE-ECRC:
CSE:

Otis-Lennon Mental Ability Test
Intelligence - Group
Psychological Corporation
1936-70
Kindergarten, 1.0-1.5, 1.6-3.9, 4-6, 7-9, 10-12
8MMY: 198
7MMY: 370
CSE-ECRC:
CSE:

Peabody Individual Achievement Test
Achievement
American Guidance Service
1970
Kindergarten - 12
8MMY: 24
7MMY: 17
CSE-ECRC:
CSE:

Peabody Picture Vocabulary Test
Intelligence - Individual
American Guidance Service
1959-65
Ages 2.5 - 18
8MMY: 222
7MMY: 417
CSE-ECRC
CSE:
HJ: 191-192

Pictorial Test of Intelligence
Intelligence - Individual
Houghton Mifflin Co.
1964
Ages 3-8
8MMY: 223
7MMY: 418
CSE-ECRC:
CSE:

Preschool Embedded Figures Test
Personality - Nonprojective
Consulting Psychologists Press, Inc.
1972
Ages 3-5
T2: 1331

Preschool Interpretation Problem-Solving Test
Personality
Myrna Shire and George Spivack
NA
Age 4-5
OJ: 565-567

Prescriptive Reading Inventory
Reading - Diagnosis
CTB/McGraw Hill
1972-77
Kindergarten.0-1.0, Kindergarten.5-2.0, 1.5-2.5, 2.0-3.5, 3.0-4.5, 4.0-6.5
8MMY: 769

Primary Academic Sentiment Scale
Reading - Readiness
Priority Innovations, Inc.
1968
Ages 4-4 to 7-3
T2: 1723
7MMY: 760
W: 147, 212
CSE-ECRC

Primary Mental Abilities Test
Multi-aptitude
Science Research Associates
1946-65
Kindergarten-1, 2-4, 4-6, 6-9, 9-12
8MMY: 488
T2: 1087
CSE-ECRC:
CSE:
HJ: 213-214

SRA Achievement Series
Achievement
Science Research Associates
1954-69
Grades 1-2, 2-4, 3-4, 4-9
7MMY: 18
T2: 731, 108, 1596, 1790, 1947, 1765

SRA Assessment Survey
Achievement
Science Research Associates
1954-75
Grades 1-2, 2-4, 4-5, 6-7, 8-9
8MMY: 1
CSE:

Santa Clara Inventory of Developmental Tasks
Readiness
Richard L. Zweig Associates, Inc.
1974
Ages Preschool, 5-5.5, 6-6.5, 7
No references

School Readiness Test
Reading - Readiness
Scholastic Testing Service, Inc.
1974-77
Kindergarten - Grade 1
8MMY: 808-9

Screening Test for Auditory Comprehension of Language
Miscellaneous - Listening
Learning Concepts
1973
Ages 3-6
8MMY: 454a
OJ: 223
CSE: 41
HJ: 235-236

Screening Test of Academic Readiness
Reading - Readiness
Priority Innovations, Inc.
1966
Ages 4-0 to 6-5
T2: 1730
7MMY: 765
CSE-ECRC:
HJ: 239-240

The Self-Concept and Motivation Inventory: What Face Would You Wear?
Personality
Person-O-Metrics, Inc.
1967-77
Age 4-kindergarten, Grades 1-3, 3-6, 7-12
8MMY: 670
OJ: 722-23
W: 249
CSE:

Short Form Test of Acacemic Aptitude
Intelligence - Group
CTB/ McGraw-Hill
1936-74
Grades 1.5-3.4, 3.5-4, 5-6, 7-9, 9-12
8MMY: 202
7MMY: 387
CSE:

Short Test of Educatioral Ability
Intelligence - Group
Science Research Associates, Inc.
1966-70
Kindergarten-1, 2-3, 4-6, 7-8, 9-12
7MMY: 382
CSE-ECRC:
CSE:

Slosson Intelligence Test
Intelligence - Individual
Slosson Educational Publications, Inc.
1961-63
Ages 2 weeks and over
8MMY: 227
7MMY: 424
CSE-ECRC:
CSE:
HJ: 243-244

Slosson Oral Reading Test
Reading - Oral
Slosson Educational Publications, Inc.
1963
Grades 108, and high school
T2: 1688
CSE:

Stanford Achievement Test
Achievement
Psychological Corporation
1923075
Grades 1.5-2.4, 2.5-3.4, 3.5-4.4, 4.5-5.4, 5.5-6.9, 7.0-9.5
8MMY: 29
7MMY: 25
CSE:
HJ:

Stanford-Binet Intelligence Scales
Intelligence - Individual
Houghton Mifflin Co.
1916-73
Ages 2 and over
8MMY: 229
7MMY: 425
CSE-ECRC:

Stanford Diagnostic Mathematics Test
Mathematics
Psychological Corporation
1976
Grades 1.5-4.5, 3.5-6.5, 5.5-8.5, 7.5-13
8MMY: 292

Stanford Diagnostic Reading Test
Reading - Diagnosis
Psychological Corporation
1966-76
Grades 1.5-3.5, 2.5-5.5, 4.5-9.5, 9-13
8MMY: 777
7MMY: 725

Stanford Early School Achievement Test
Achievement
Psychological Corporation
1969-71
Kindergarten-1.1, 1.1-1.8
8MMY: 30
7MMY: 28
CSE-ECRC:
CSE:
HJ: 249-250

Steinbach Test of Reading Readiness
Reading - Readiness
Scholastic Testing Service Inc.
1965-66
Kindergarten - Grade 1
T2: 1732
CSE-ECRC:

Templin-Darley Tests of Articulation
Speech & Hearing - Speech
Bureau of Educational Research and Service
1960-69
Ages 3 and over
T2: 2095
7MMY: 972
CSE-ECRC:
HJ: 253-254

Test of Language Development
Speech & Hearing - Speech
Empiric Press
1977
Ages 4-0 to 8-11
8MMY: 978

Test of Nonverbal Auditory Discrimination
Speech & Hearing - Hearing
Follett Publishing Co
1968-75
Kindergarten - 3
8MMY: 950
QJ: 947

Tests of Basic Experiences
Achievement
CTB/McGraw Hill
1970-75
Prekindergarten - Kindergarten (Level K), Kindergarten - Grade 1, (Level L)
8MMY: 34
7MMY: 33
CSE-ECRC:
CSE:
HJ: 257-258

Valett Developmental Survey of Basic Learning Abilities
Reading - Readiness
Consulting Psychologists Press, Inc.
1966
Ages 2-7
T2: 991
7MMY: 767
CSE-ECRC:
CSE:
HJ: 267-268

Vineland Social Maturity Scale
Personality
American Guidance Service
1935-65
Birth to maturity
8MMY: 703
W: 301-302
CSE-ECRC:
CSE:
HJ: 273-274

Walker Readiness Test for Disadvantaged Preschool Children
Readiness
Wanda H. Walker
Age 4 to 6 years
OJ: 154-155

Wechsler Intelligence Scale for Children
Intelligence - Individual
Psychological Corporation
1949-74
Ages 5-16
8MMY: 232
7MMY: 431
CSE-ECRC:
CSE:

Wechsler Preschool and Primary School Intelligence Test
Intelligence - Individual
Psychological Corporation
1949-67
Ages 4-5.5
8MMY: 234
7MMY: 434
CSE-ECRC:
CSE:

Wide Range Achievement Test
Achievement
Guidance Associates of Delaware, Inc.
1940-76
Ages 5-11, 12 and over
8MMY: 37
7MMY: 36
CSE-ECRC:
CSE:
HJ: 279-280

Woodcock Reading Mastery Test
Reading - Diagnosis
American Guidance Service
1972-73
Kindergarten - 12
8MMY: 779
CSE:

EARLY CHILDHOOD TEST INSTRUMENT PUBLISHERS

Addison-Wesley Publishing Co., Inc.
2725 Sand Hill Road
Menlo Park, California 94025

American Guidance Service, Inc.
Publishers' Building
Circle Pines, Minnesota 55014

American Orthopsychiatric Association, Inc.
1775 Broadway, New York, New York 10019

Bobbs-Merrill Co., Inc. (The)
4300 West 62nd Street
Indianapolis, Indiana 46268

Bureau of Educational Research and Service
University of Iowa
Iowa City, Iowa 52242

CTB/McGraw Hill
Del Monte Research Park
Monterey, California 93940

Consulting Psychologists Press, Inc.
577 College Avenue
Palo Alto, California 94306

Cooperative Tests and Services
c/o Addison-Wesley Publishing Co., Inc.
2725 Sand Hill Road
Menlo Park, California 94025

Crippled Children and Adults of Rhode Island, Inc.
Meeting Street School
667 Waterman Avenue
East Providence, Rhode Island 02914

Educational Development Corporation
P.O. Box 45663
Tulsa, Oklahoma 74145

Educational Testing Service
Princeton, N.J. 08540

Empiric Press
333 Perry Brooks Building
Austin, Texas 78701

Follett Publishing Co.
1010 West Washington Boulevard
Chicago, Illinois 60607

Guidance Associates of Delaware, Inc.
1526 Gilpin Avenue
Wilmington, Delaware 19806

Harcourt Brace Jovanovich, Inc.
757 Third Avenue
New York, New York 10017

Houghton Mifflin Company
1 Beacon Street
Boston, Massachusetts 02107

Learning Concepts
2501 North Lamar
Austin, Texas 78705

Person-O-Metrics, Inc.
20504 Williamsburg Road
Dearborn Heights, Michigan 48127

Priority Innovations, Inc.
P.O. Box 792
Skokie, Illinois 60076

Psychological Corporation (The)
757 Third Avenue
New York, New York 10017

Research Concepts
1368 East Airport Road
Muskegon, Michigan 49444

Richard Zweig Associates, Inc.
20800 Beach Boulevard
Hunnington Beach, California 92648

Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, Illinois 60106

Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606

M. Shufe and G. Spivack
Community Mental Health
Mental Retardation Center
Department of Mental Health Sciences
Hahneman Medical College and Hospital
Philadelphia, Pennsylvania 19102

Slosson Educational Publications, Inc.
140 Pine Street
East Aurora, New York 14052

Teacher's College Press
1234 Amsterdam Avenue
New York, New York 10027

University of Illinois Press
Urbana, Illinois 61801

Virginia Research Associates, Ltd.
P.O. Box 5501
Charlottesville, Virginia 22902

Wanda Walker
Northwest Missouri State College
Morgsville, Missouri 64468

Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, California 90025

Appendix 3

ANNOTATIONS ON EARLY CHILDHOOD INSTRUMENTS

This appendix provides brief annotations concerning five additional instruments:

Animal Crackers
CTBS Readiness Test
Santa Clara Inventory of Developmental Tasks
Preschool Inventory
Wechsler Preschool and Primary Scale of Intelligence

These annotations are provided simply to illustrate the type of information useful in initially screening instruments for possible use. The fact that particular instruments are listed here should not be interpreted as an endorsement.

Annotation Form

TITLE: Animal Crackers: A Test of Motivation to Achieve FORMS:
COPYRIGHT: 1973

AUTHOR: Dorothy C. Adkins & Bonnie L. Ballif

PUBLISHER: CTB/McGraw-Hill

SOURCE: CTB/McGraw-Hill, Del Monte Research Park,
Monterey, CA 93940

PRICE AS OF 1980: $18.60

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: personality test

INTENDED USE: to assess achievement motivation, how the child feels about himself in the school situation and whether or not learning is important to him

INTENDED POPULATION: preschool, kindergarten, and first grade

ITEM FORMAT: objective-projective technique (the child chooses between alternative behaviors or attitudes, described orally). Each item consists of an illustration of two identical animals and two oral descriptions. The child is told that he has his "own" animals which look like the others but behave as he does. As the examiner points to each animal in turn and describes it, the child identifies his own animal.

CONTENT: School enjoyment
Self-confidence
Purposiveness
Instrumental activity
Self-evaluation

II. REFERENCES

1. Adkins, Dorothy C. & B.L. Ballif. Examiner's Manual, Research Edition: Animal Crackers, A Test of Motivation to Achieve. Monterey, CA: CTB/McGraw-Hill, 1973.

2. Weintraub, S. Review in Buros' Eighth Mental Measurements Yearbook, pp. 693-694.

Annotation Form

TITLE: CTBS Readiness Test

FORMS: Level A, Form S
COPYRIGHT: 1977

AUTHOR: CTB/McGraw-Hill

PUBLISHER: CTB/McGraw-Hill

SOURCE: CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA 93940

PRICE AS OF 1980: $5.90, specimen set

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: readiness test

INTENDED USE: "to help kindergarten and first grade teachers and supervisors determine if their students have the skills necessary for beginning reading"; to diagnose strengths and needs in particular skill areas; to predict success in reading

INTENDED POPULATION: Grades K.0 - 1.3

ITEM FORMAT: multiple choice (children fill in circle corresponding to correct choice)

CONTENT: letter forms
letter names
listening for information
letter sounds
visual discrimination
sound matching
language
mathematics

II. REFERENCES

CTBS Readiness Test: User's Handbook for the Reading Readiness Report of Skill Mastery. Monterey, CA: CTB/McGraw-Hill, 1977.

CTBS Readiness TEST: Examiner's Manual. Monterey, CA: CTB/McGraw-Hill, 1977.

CTBS Readiness Test: Test Book. Monterey, CA: CTB/McGraw-Hill, 1977.

Findley, W. Review of Comprehensive Test of Basic Skills, Expanded Edition, Buros' Eighth Mental Measurements Yearbook, pp. 40-43.

Nitko, A. Review of Comprehensive Test of Basic Skills, Expanded Edition, Buros' Eighth Mental Measurements Yearbook, pp. 43-45.

Annotation Form

TITLE: Santa Clara Inventory of Developmental Tasks FORMS:
COPYRIGHT: 1974

AUTHOR: William L. Gainer (senior editor), et al.

PUBLISHER: Richard L. Zweig Associates, Inc.

SOURCE: Richard L. Zweig Associates, Inc., 20800 Beach Boulevard, Huntington Beach, California 92648

PRICE AS OF 1980 : $59.95

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: mastery test

INTENDED USE: as a screening device to assess children's developmental skills and for learning disabilities and to guide on-going instruction as part of a "failure-prevention program"

INTENDED POPULATION: preschool through age 7

ITEM FORMAT: children give motor, oral, or written responses to oral instructions

CONTENT: 60 tasks sequenced by chronological age and arranged by eight skill areas: motor coordination, visual motor performance, visual perception, visual memory, auditory perception, auditory memory, language development, and conceptual development

---

II. REFERENCES

Gainer, William L. (senior editor) et al. Observation Guide for the Santa Clara Inventory of Development Tasks. Huntington Beach, CA: Richard L. Zweig Associates, Inc., 1974.

Annotation Form

TITLE: Preschool Inventory, Revised Edition

FORMS:
COPYRIGHT: 1970

AUTHOR: Dr. Betty M. Caldwell

PUBLISHER: Addison-Wesley Testing Service

SOURCE: Addison-Wesley Testing Service, South Street, Reading, MA 01867

PRICE AS OF 1979: $3.00, specimen set

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: individual intelligence test

INTENDED USE: to assess knowledge and skills children bring with them to primary education; "to highlight the degree of disadvantage which a child from a deprived background has at the time of entering school so that any observed deficiencies might be reduced or eliminated"; and "to demonstrate changes associated with educational intervention."

INTENDED POPULATION: children ages 3 to 6

ITEM FORMAT: children give oral or motor responses to questions

CONTENT: basic information and vocabulary
number concepts and ordination
concepts of size, shape, motion, and color (Concepts I)
concepts of time, object loss, and social functions (Concepts II)
visual-motor performance
following instructions
independence and self-help

II. REFERENCES

Caldwell, B.M. Preschool Inventory, Revised Edition - 1970: Handbook. Reading, MA: Addison-Wesley Publishing Company, 1970.

Caldwell, B.M. Preschool Inventory, Revised Edition - 1970: Directions for Administering and Scoring. Reading, MA: Addison-Wesley Publishing Company, 1970.

French, J. Review of Buros, O.K. Seventh Mental Measurements Yearbook. pp. 730-732.

Annotation Form

TITLE: Wechsler Preschool and Primary Scale of Intelligence (WPPSI)

FORMS:
COPYRIGHT: 1963, 1967

AUTHOR: David Wechsler

PUBLISHER: The Psychological Corporation

SOURCE: The Psychological Corporation, 304 East 45th St., New York, NY 10017

PRICE AS OF January, 1979: $34.75 for the basic set

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: individual intelligence test

INTENDED USE: to appraise the abilities of the preschool child; it consists of a battery of subtests which, treated separately, may be considered as measuring a different ability and which, combined, provide a measure of global intellectual capacity; for diagnosis.

INTENDED POPULATION: children ages 4 through 6 1/2 years

ITEM FORMAT: children give motor, oral, or written responses to oral instructions

CONTENT: 11 tests:

Verbal Tests	Performance Tests
Information	Animal House
Vocabulary	Picture Completion
Arithmetic	Mazes
Similarities	Geometric Design
Comprehension	Block Design
Sentences (Supplementary Task)	

II. REFERENCES

Buros, O.K. (Ed.) Eighth Mental Measurements Yearbook. New Jersey: Gryphon Press, 1978.

Wechsler, David. Manual for the Wechsler Preschool & Primary Scale of Intelligence. New York: The Psychological Corporation, 1967.

Appendix 4

REVIEWS OF SELECTED EARLY CHILDHOOD INSTRUMENTS

This appendix provides information concerning seven early childhood instruments, in terms of the outline given in Table 1 above. The seven instruments reviewed here are, in alphabetical order:

Boehm Test of Basic Concepts
CIRCUS
McCarthy Scales of Children's Abilities
Metropolitan Readiness Tests
Peabody Picture Vocabulary Test
Stanford Early School Achievement Test
Test of Basic Experiences 2

These reviews are provided to illustrate the type of information useful in selecting among candidate instruments. The fact that these particular instruments are reviewed should not be interpreted as an endorsement. For any particular purpose, other instruments should also be considered.

Review Form

TITLE: Boehm Test of Basic Concepts, 1971 Edition

FORMS: A and B
COPYRIGHT: 1967, 1969, 1970, 1971

AUTHOR: Ann E. Boehm

PUBLISHER: The Psychological Corporation

SOURCE: The Psychological Corporation, 757 Third Avenue, New York, NY 10017

PRICE AS OF 1979: $1.60 (Specimen set); $7.35 (test booklets, each form, package of 30 each of booklets 1 and 2)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: group intelligence test

INTENDED USE: to measure children's knowledge of concepts that are often assumed to be familiar to them when they enter school and that are considered necessary for early school achievement; for individual remediation, group instruction, or both.

INTENDED POPULATION: children in grades K-2

ITEM FORMAT: 50 pictorial multiple choice items, read aloud by the teacher; three or four items per page. Children mark their choices with x's and make corrections by encircling or erasing the error.

CONTENT: 50 widely used terms that involve simple relationships between/among objects

---

II. THEORY, CONSTRUCTION, AND DEVELOPMENT

Form A of the BTBC was developed in 1969 to assess mastery of certain basic concepts. The rationale for the test is that children are assumed to know these concepts when they enter school and that success in school depends on that knowledge.

The specific concepts were chosen after a review of a "comprehensive selection" of early childhood curriculum materials in the areas of reading, arithmetic, and science. Terms were chosen that met the following criteria: (1) occurred frequently, (2) were defined inadequately (if at all), and (3) represented relatively abstract basic concepts or ideas. Pictorial multiple choice questions were drawn up to test these concepts and were tested twice in October 1967, first to make an initial selection of concepts, and second to make any necessary revisions or adjustments. Badly drawn or ambiguous items were eliminated, along with those answered correctly by 98% or more (thus, concept unfamiliarity became a criterion for final selection). Form B items were written to parallel Form A, so that corresponding items measure knowledge of the same basic concept.

The first item analysis sample involved 267 children in kindergarten, 306 in Grade 1, 264 in Grade 2, and 297 in Grade 3, from four schools in a

city of about 75,000 in southeastern New York State. One school was composed of predominantly high SES children, another of low SES children, and the other two were mixed.

The second group tested comprised 671 children in kindergarten, 823 in Grade 1, 710 in Grade 2, from six schools in an eastern city of about two million. Third graders were not included, since they had found the items so easy in the first tryout that the test was determined to have limited usefulness for average children at this grade level.

III. PRACTICAL CONSIDERATIONS

MATERIALS REQUIRED: all provided in test package (Manual, test booklets--two for each form--and a class record sheet for each form)

TYPE OF ADMINISTRATION (group or individual): group

ADMINISTRATION TIME: untimed test; each booklet takes about 15 to 20 minutes with kindergartners

SETTING: in groups of 8-12 children; with larger groups, the use of proctors or teacher's aides is advised

DIRECTIONS: the Manual prints test directions in boldface, to be read aloud, with key phrases for each item read twice and emphasis placed on italicized words

SAMPLE QUESTIONS: three sample questions are part of each booklet in each form

SCORING PROCEDURE: tests are scored by hand

LANGUAGE OF ADMINISTRATION: English or Spanish directions are available

TRAINING NEEDED TO ADMINISTER: no special training is required

EARLY CHILDHOOD CONSIDERATIONS: the BTBC is a group test, with three or four items per page. The test is fairly short, and practice questions are provided.

IV. TECHNICAL INFORMATION

NORMING POPULATION: 9,737 children in grades kindergarten through two in 16 cities, for beginning-year norms; 2,669 K-2 children from five cities, for the mid-year norms

NORMING SAMPLE METHOD: classrooms were chosen from schools that, in the judgment of school officials, represented a wide range of SES groups

NORMING MIDPOINTS: beginning-of-year (September - October) and mid-year (December - January)

SCALES AND SCORES: percent of children passing each item and raw score to percentile conversions (percentiles listed are the approximate midpoint of a range of percentiles that are five percentiles wide, except above 95 and below 5, where the "bands" of percentiles are smaller)

VALIDITY: the manual says that evidence for content validity is the fact that the items were selected from relevant curriculum materials. No other forms of validity are discussed.

RELIABILITY: split-half reliability coefficients and standard errors of measurement are presented by SES level within grade for both forms at the mid-year test point. The coefficients for Form A range from .68 to .90, and for Form B from .12 to .94 (the author dismisses the .12, which was obtained for the Grade 2 high SES group, because the children in the group represent the upper end of the ability groups for which the test was intended). Alternate form reliability coefficients range from .55 to .92, with a median of .76.

SPECIAL NOTE: the norms for Form B are estimates based on the results of equivalence study of Forms A and B. First, the forms were equated using the equipercentile method. Although the forms did not correspond perfectly, scores on Form B were never more than 1.5 points higher than comparable scores on Form A. Thus, the forms are considered equivalent. 1,292 children in grades K-2, from several SES groups in three cities, were tested in December-January, 1970-71. Then the equivalence of items was established, by dividing the data from the two testings into nine cells (three grades times three SES levels), determining the difficulty level for each item, and calculating regression equations to predict the difficulty of an item on Form b for a given group based on its score on the corresponding Form A items. These results were then used to establish the norms for Form B.

V. OUTSIDE REVIEWS

Excerpts of a review by Theodore A. Dahl appear in the Eighth Mental Measurements Yearbook (p. 253). The Seventh Mental Measurements Yearbook contains two complete reviews, by Boyd R. McCandless (p. 626) and Charles D. Smock (p. 626), and four excerpted reviews, by Victor H. Noll (p. 627), Barton B. Proger (p. 628), George Lawlor (p. 628), and Frank S. Freeman (p. 629). Reviewers agreed that the rationale of the test was noteworthy, overall booklet design was good, and administrative and scoring procedures were clear. Beyond that, the reviews present contradictory opinions, notably regarding test validity and interpretation of results. Smock called the content validity adequate, and he praised the discussion of interpretation. McCandless spoke of face validity, which he said was convincing, but noted the lack of data on other forms of validity. Proger felt information on predictive, construct, and concurrent validity would have been helpful, along with a review of the literature underlying the test's development. Noll observed, further, that the test's validity suffered because the manual presented no information about the curriculum materials examined, nor on

the relationship between mastery of the concepts measured and school achievement; moreover, the criteria for selecting items were imprecise. Freeman elaborated on the validity issue by noting the lack of evidence that the items were comprehensive and representative regarding the curriculum materials, and by pointing out that there are undoubtedly other concepts children should know that are not tested here. The strongest criticism on this subject was Dahl's. He argued that the BTBC's purpose is really to function as a criterion-referenced test, with mastery of 50 concepts as the objective. Yet the manual's technical discussion treats it as a norm-referenced test, while adding the confusing statement that the main purpose of the test is such that norms are not necessary. If the test measures 50 specific concepts, Dahl maintains, then it should be considered 50 separate one-item tests. In this case, however, both reliability and validity would be highly questionable.

Several other comments were made by only one or two reviewers. Lawlor, McCandless, and Noll noted that the test is too easy for Grade 2 students and for Grade 1 children from middle and upper SES levels. McCandless also questioned the recommended use of the test with groups of eight to 12 children without extensive use of proctors. Freeman questioned the non-representative standardization population; and Lawlor felt it should have included nursery school subjects. However, the test won praise by Dahl for its interest level, by Proger for its encouragement of program evaluation, discussion of confidence intervals and standard errors of measurement, and implications for disadvantaged and handicapped children.

One other outside review, by CSE-ECRC, gave the BTBC ratings of fair in terms of measurement validity and normed technical excellence, and good in terms of examinee appropriateness and administrative usability.

VI. COMMENTS

GENERAL INFORMATION: as a test of understanding of basic concepts, the BTBC has a limitation, which the manual acknowledges in a footnote: a child may know a concept when applied to physical objects, but not on the printed page. For example, a child may lack the artistic perspective necessary to determine "which boat is farthest from the shore," yet understand the concept "farthest." The author claims this limitation is justified, since school deals with printed materials.

Although the test covers "basic concepts," one wonders whether these are comprehensive. Even if necessary for school achievement (and no evidence is presented to support this), are they sufficient? In other words, will mastery of all 50 concepts assure success in school or even contribute to it greatly?

THEORY, CONSTRUCTION, AND DEVELOPMENT: the test should be commended for seeking to articulate what is often unjustifiably assumed--children's knowledge of basic concepts at the time of school entry--and for attempting to provide information to help teach such concepts to children who do not know them. However, discussion in the manual of the theoretical basis for the choices of items is inadequate. The curriculum materials used are not identified, and, as noted above, no evidence is given to show the rela-

tionship between mastery of the 50 concepts and school achievement (although the relationship is intuitively plausible).

The BTBC is designed for grades K through 2 (and norms are provided for those levels), and it is said to be adaptable for preschool (although no norms are available). However, the norm tables suggest the test is too easy for most second graders and for all but low SES first graders.

Parts of the discussion on interpreting and using test results are noteworthy. In particular, the manual makes the unusual and laudable suggestion that the teacher ask the children why they answered the questions as they did. By understanding more clearly how the child perceives the concept (or the test item), the teacher can better plan remedial activities.

PRACTICAL CONSIDERATIONS: the BTBC manual clearly describes administrative and scoring procedures, and norms tables are easy to read. However, teachers would be well advised to administer the sample questions some time before the test, in order to have time to work with children on the proper way to mark answers. Three sample questions given immediately before the test items seem insufficient for young children who may be unfamiliar with test-taking procedures.

Although the test length is appropriate for children of ages 5-6, use of the BTBC with this age level may be difficult with the recommended group size of eight to 12. Given the kind of attention that is necessary to assure that all of the children are working on the right item and understand what the task requires them to do, this number seems large.

Finally, the number of items per page may create a distraction for children at this young age. Each page has three or four items, and if adjacent pages are open, they form a larger set of six to eight groups of pictures. Although each page has a page finder, it is quite possible that children's attention will wander.

TECHNICAL INFORMATION: norms are not nationally representative, although the author contends that they do not need to be. Discussion of their use is not clear. The raw score conversions are noteworthy in that they translate scores to a range of percentiles, rather than to a single percentile.

Lack of evidence of test validity is a major weakness of the BTBC. Content validity is assumed, on the basis that items were chosen from the curricular materials to which they are supposed to relate, but this relationship is not documented. The manual fails to discuss predictive validity, although concepts were chosen that are believed to relate to school success. Similarly, construct and concurrent validity are not mentioned.

VII. REFERENCES

Boehm, A.E. Boehm Test of Basic Concepts, 1971 Edition. New York: The Psychological Corporation, 1971.

Forms A & B

Class Record Form, Forms A & B

Booklet 1, Forms A & B

Booklet 2, Forms A & B

Buros, O.K. (Ed.) Seventh Mental Measurements Yearbook. New Jersey: Gryphon Press, 1972.

Buros, O.K. (Ed.) Eighth Mental Measurements Yearbook. New Jersey: Gryphon Press, 1978.

Hoepfner, R., Stern, C., and Nummedal, S. (Eds.) CSE-ECRC Preschool/ Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1971.

Review Form

TITLE: CIRCUS

FORMS: Levels A & B
COPYRIGHT: 1976

AUTHOR:

PUBLISHER: Addison-Wesley Publishing Company

SOURCE: Addison-Wesley Testing Service, South Street, Reading, MA 01867

PRICE AS OF 1980: $10.50 specimen set, each level

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: group administered diagnostic instrument

INTENDED USE: For classroom use, to provide an array of instruments that would be useful in diagnosing the instructional needs of individual children and in monitoring and evaluating early education programs.

INTENDED POPULATION: children in nursery school, kindergarten, and first grade. Level A - nursery to beginning kindergarten; Level B - end of kindergarten to beginning of grade one.

ITEM FORMAT: multiple choice (children mark the picture they choose as an answer); Level A presents no more than two items per page; Level B, no more than three items per page; two of the measures (Activities Inventory and Educational Environment Questionnaire) are completed by the teacher.

CONTENT: Circus consists of 16 diagnostic measures and an Educational Environment Questionnaire. Users are expected to select a single measure or combination of measures, administered in any order. The content areas and individual measures are:

	Level A	Level B
Practice materials	x	x
Language and reading		
How words sound	x	
What words mean	x	
How words work	x	
Listen to the story	x	x
Say and tell	x	x
Word puzzles		x
Perception		
Look-alikes	x	x
Copy what you see	x	

[x]Indicates the measures in Level A and the comparable measures in Level B

	Level A	Level B
Perception (con't)		
Finding letters and Numbers	*	*
Noises	x	
Mathematics		
How much and how many	x	x
Information processing and experience		
Do you know . . . ?	x	x
See and remember	x	x
Think it through	x	x
Divergent production		
Make a tree	*	*
Attitudes and interests		
Things I like		x
Activities inventory	x	x
Teacher-Program measure		
Educational Environment Questionnaire	x	x

[x] Indicates the measures in Level A and the comparable measures in Level B

[*] Indicates the same measure was normed with nursery schools, kindergarten, and first graders

---

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST

The developers examined measurement approaches, measurement content, and implications for classroom use in ETS research activities on disadvantaged young children , Sesame Street, school readiness, and social competency in young children. The eventual array of instruments was chosen "on the basis of whether the domain was measurable in a straightforward way, whether it had direct implications for the education of children, and on the basis of cost, time, psychometric confidence, and tryout and rejection of prototype items and measures." [no further information given]

III. PRACTICAL CONSIDERATIONS

MATERIALS REQUIRED: most measures provide all the necessary materials (test booklet and directions to teacher). "Say-Tell" calls for two pennies and a watch as well.

TYPE OF ADMINISTRATION (group or individual): group, up to 10 children (except for the "Say and Tell" test, which is administered individually)

ADMINISTRATION TIME: most measures untimed; children proceed at own pace

SETTING: an area free of distractions, quiet, well-lit, apart from children who may be tested later

DIRECTIONS: each measure comes with its own user's guide (approximately 25 pages each), explaining the rationale for and interpretation and scoring of the measure. The directions are clearly worded but with some tests the user must read several sources (the Technical Manual and the individual instrument, for example).

SAMPLE QUESTIONS: practice materials are included to prepare children for the mechanical test-taking tasks. In addition, most measures have one or two sample questions.

SCORING PROCEDURE: hand scoring or machine scoring by ETS; in addition to numerical scores, all of the measures yield "Sentence Reports," or verbal descriptions of the child's performance

LANGUAGE OF ADMINISTRATION: English (a Spanish version, CIRCO, is being developed)

TRAINING NEEDED TO ADMINISTER: none; test can be administered by teachers' aides as well as teachers

---

IV. TECHNICAL INFORMATION

NORMING POPULATION: level A was normed on 1,979 kindergarten and 1,006 nursery school children attending 145 preprimary education centers. Level B was normed on 6, 394 children in 286 first grade classes. Level B testing did not include handicapped and special education classes.

NORMING SAMPLE METHOD: level A's norm sample "includes children in four regions of the country from various city sizes and socioeconomic and ethnic backgrounds." Selection procedure is not specified. Some measures were not administered to all children, so that "sample data for a particular test may not accurately represent the entire population of preprimary children or a particular subgroup of children." Level B's sample includes children randomly selected from 17 strata based on data obtained from the U.S. Department of Commerce County and City Data Book, 1972 on region of the country, urbanization, and SES. Since this design produced an equal number of children in each type of county in each region of the country, the strata were weighted so that the weighted sample corresponded to the total population with respect to the variables on which the sample was stratified. Each child was administered only five of the Circus instruments.

NORMING MIDPOINTS: level A was normed in January (nursery school) and October (kindergarten and first grade)

SCALES AND SCORES: sentence reports, some percentiles, "percent obtaining scores" and r-biserials. Also, some Growth Expectancy Tables provided

VALIDITY: correlations are provided between CIRCUS A and CIRCUS B measures. These range from .34 to .74. Correlations between teacher judgment and CIRCUS A range from -.14 for portions of one measure to .56, and between teacher judgment and CIRCUS B from .16 to .58.

RELIABILITY: Alpha Reliability (internal consistency) reported for all of the measures and subscales; ranging from .68 to .92 (nursery school), .71 to .87 (kindergarten) and .75 to .93 (first grade). Standard errors of measurement are also reported.

V. OUTSIDE REVIEWS

CIRCUS has not yet been reviewed in Buros' MMYs. For brief comments on this instrument, see Johnson, H. Preschool Test Descriptions, Springfield, Illinois, 1979, pp. 85-86.

VI. COMMENTS

THEORY, CONSTRUCTION, AND DEVELOPMENT: Very little explanation is given concerning the development of the CIRCUS instruments, aside from the observation that such an array of instruments would be useful. Item tryout and analysis, though alluded to, are not described.

PRACTICAL CONSIDERATIONS: The CIRCUS instruments, individually, appear to be straightforward to administer. Collectively, however, when ordered as a specimen set, they appear complex and confusing. Each of the 17 measures (for each level) comes with its own user's guide, as well as a test booklet and local scoring class performance record. In addition, a 108-page manual accompanies the measures and provides an overview and additional administrative and technical information. All together there is a lot of information for a teacher to plow through in order to select the measures best-suited to his/her needs. However, the notion of a test package that encourages flexibility in the choice and order of tests to be administered is appealing in its adaptability to individual user's needs. The test scores include an extensive "sentence report", which corresponds to the intended use of diagnosing individual instructional needs.

CIRCUS should be appealing to young children because of its simple, clear pictures, its limited number of items per page, its brevity, and its circus theme. However, it is a group test.

TECHNICAL INFORMATION: Some of the norming information for Level A is insufficient. For example, sample selection procedures are not clearly specified. In addition, it should be noted that the whole test was not standardized on the whole norming

population, but that only 4 (Level A) and 5 (Level B) instruments were given to each child. Additional validity evidence, for example concerning predictive validity would be helpful.

VIII. REFERENCES

Circus: Manual and Technical Report, Teacher's Edition. Reading, MA: Addison-Wesley Publishing Company, 1976.

The 17 individual measures are also available with separate user's guides.

Review Form

TITLE: McCarthy Scales of Children's Abilities

FORMS:

COPYRIGHT: 1970, 1972

AUTHOR: Dorothea McCarthy

PUBLISHER: The Psychological Corporation

SOURCE: The Psychological Corporation, 757 Third Avenue, New York, NY 10017

PRICE AS OF 1979: $70.00 (complete kit; no specimen set available)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: individual intelligence test

INTENDED USE: to determine young children's general intellectual level as well as their strengths and weaknesses in important abilities

INTENDED POPULATION: ages 2 1/2 through 8 1/2, both sexes, from various ethnic, regional, and socioeconomic groups

ITEM FORMAT: children give motor or oral responses to oral instructions; examiner observations of laterality also provide information.

CONTENT: 18 tests grouped into 6 scales:

Verbal (V)
Perceptual-Performance (P)
Quantitative (Q)
General Cognitive (GC)
Memory
Motor

The Verbal, Perceptual-Performance, and Quantitative Scales are non-overlapping, such that V + P + Q = GC. All of the Memory and some of the Motor Scales overlap with the General Cognitive Scales.

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST:

The test was developed in response to the recognition of the importance of a child's early years and the role they play in his cognitive development; and in response to the awareness of a lack of appropriate tools for assessing young and retarded children.

"A wide variety of tasks involving language, numerical concepts, motor coordination, and other skills generally considered to reflect cognitive and motor ability, were selected for the early tryouts . . . " with mentally retarded and normal children to explore the upper and lower ranges and the tasks' ability to discriminate among successive age levels. Several (unspecified) tasks were redesigned and tested on "more rigorously defined populations," before further refinements were made. - The choice of content and the grouping of tests into scales were based on the author's teaching and clinical experience and on the results of factor analyses.

III. PRACTICAL CONSIDERATIONS:

MATERIALS REQUIRED: Supplied in the testing kit - drawing booklets; card book; 12 1-inch cubes of the same color; 6 puzzles; xylophone and mallet; tape; rubber ball; beanbag; target; tube; 2 5x8 inch pieces of cardboard; set of 12 blocks, each two shapes, two each in three colors, and two sizes per color. Materials supplied by examiner - stopwatch; scotch tape, non-transparent tape; soft paper (with scotch tape or rubber band); two no. 2 black pencils, one 4-6 inches with no eraser; sheets of unlined 8 1/2x11 inch paper; extra toy or book to help establish rapport.

TYPE OF ADMINISTRATION (Group or Individual): Individual

ADMINISTRATION TIME: estimated at 45 to 50 minutes for children under 5; at 60 perhaps up to 75 minutes for older children.

SETTING: testing room should contain an adult-sized and a child-sized chair, and a table surface that is below the child's line of vision when seated. The room should be uncarpeted and have 10 feet of clear space (or a nearby quiet hallway).

DIRECTIONS: instructions to be read by the teacher are given in red. Instructions to the teacher and scoring standards are in black. Pictures are provided to illustrate the correct arrangement of materials.

SAMPLE QUESTIONS: none

SCORING PROCEDURES: hand-scoring

LANGUAGE OF ADMINISTRATION: English

TRAINING NEEDED TO ADMINISTER: the test battery and scoring are complicated and require practice and familiarity

EARLY CHILDHOOD CONSIDERATIONS: The test is administered individually, alone if possible. The examiner is advised to avoid the word "test," unless directly asked. To establish rapport, the examiner tells the child he is going to show the child some new puzzles and blocks, ask him some questions, and play some games with him. The 18 tests are arranged so that the child gets a "break" in the middle, when the tests of motor skills are administered. In addition, the first tests are relatively easy and do not require verbal responses, and thus may serve as "icebreakers."

IV. TECHNICAL INFORMATION

NORMING POPULATION: 1,032 "normal" children, from 100 to 106 in each of the 10 age levels

NORMING SAMPLE METHOD: A nationwide stratified sample was chosen (October 1970 through September 1971), based on age, sex, color, geographic region, and father's occupation, in proportion to 1970 U.S. census estimates. Also, urban vs. rural residence was used as an "informal selection variable." Only one child per family was tested except for three "hard-to-fill" slots, and only English-speaking children were included.

NORMING MIDPOINTS: Ten age groups: 2 1/2, 3, 3 1/2, 4, 4 1/2, 5, 5 1/2, 6 1/2, 7 1/2, 8 1/2. All cases were tested within 30 days of their birthday or half-birthday.

SCALES AND SCORES: For each of the six scales, raw score converted to scale score, called an Index, according to chronological age. The General Cognitive Index provides a general level of intellectual functioning, with a mean of 100 and a standard deviation of 16. The six scales were weighted based on the author's judgment of the relative importance of the individual tests and on the size of the standard deviations of the raw scores on the test for the normative group.

Percentile ranks given for scale indexes. Estimated mental ages are presented for clinical use in an appendix. Means and standard deviations but not norms provided for each of the 18 tests.

VALIDITY: Concurrent - based on 35 white first graders in New York City parochial schools; correlation between MSCA General Cognitive and WPPSI Full Scale IQ was .71, between MSCA General Cognitive and Stanford-Binet (Form L-M) IQ, .81. Except for MSCA Motor Scales, other correlations range from .27 to .61 for WPPSI, and .4 to .70 for Stanford-Binet.

Predictive - 31 of the above 35 children were tested four months later with the Metropolitan Achievement Tests, 1970 edition, Primary I, Form F. The correlation between MSCA General Cognitive and MAT Total Raw Score was .49.

RELIABILITY: Internal Consistency (Split-Half, with Spearman-Brown correction) - based on 100 to 106 children in each of the 10 age groups. The reliability coefficient for the General Cognitive Scale ranged from .90 to .96 for the 10 age groups. Other scales averaged from .79 to .88.

Stability (Test-Retest) - based on 125 children at six age levels who were retested after approximately one month. The reliability coefficient for the General Cognitive Scale ranged from .89 to .91, and for the other scales, from .69 to .89.

V. OUTSIDE REVIEWS:

The Eighth Mental Measurements Yearbook contains reviews by three people (Joan Hunt, pp. 309-311; Jerome Sattler, pp. 311-313; and Arthur B. Silverstein, pp. 313-314) and excerpts from four sets of reviewers (Everett

E. Davis, p. 314; Linda Hufano & Ralph Hoepfner, pp. 314-315; R.B. Ammons & C.G. Ammons, p. 315; and Alan Krichev, pp. 315-316).

On the plus side, Hunt praises the standardization procedures and design and notes the clear instructions. She maintains that the diagnostic potential of individual scores gives the MSCA an advantage over the Binet and the WPPSI. Sattler speaks of the overall excellence of the test, which he calls very promising, well standardized, and psychometrically sound. Davis & Krichev agree that the technical aspects are beyond reproach, while Ammons & Ammons and Krichev commend the clarity and thoroughness of the directions.

In contrast to the technical excellence mentioned, Sattler, Silverstein, and Ammons & Ammons point to the limited validity evidence. Praising the test overall, Davis claims the major problem concerns the construct of mental ability. Both he and Hufano & Hoepfner note the extensive training required and the complex and sometimes subjective scoring procedures, which Hunt felt had potential for miscalculations. Ammons & Ammons found the time required was considerably longer than the manual indicates, enough perhaps to suggest two testing sessions to maintain interest. While Hunt praises the MSCA's diagnostic potential, Hufano & Hoepfner argue that the test is not appropriate for that purpose, given the high test intercorrelations. And they note that the norms are not based on the exceptional populations with whom the test is to be used. Finally, Silverstein says it's too early to compare the MSCA with its competitors, and Sattler suggests that though the MSCA is promising, more information is needed.

VI. COMMENTS:

THEORY, CONSTRUCTION, AND DEVELOPMENT: Although the author prefers to avoid the term "IQ," it is not clear how the General Cognitive Index, with its mean of 100 and standard deviation of 16, differs from a traditional IQ score.

PRACTICAL CONSIDERATIONS: The test requires considerable familiarity and training. The manual contains 209 pages. Administration involves many materials and arranging of those materials and may be somewhat cumbersome. However, the test is conveniently packaged, and the manual is very clear and explicit. The test may be long for young children, although the tasks are arranged to capture and maintain their attention, including a built-in "intermission." Also, the tasks should be appealing and are presented so as to be nonthreatening. The materials are well constructed and easy to manipulate.

TECHNICAL INFORMATION: The reliability information is solid, but validity information is limited. It's based on a small, non-representative group, and predictive validity is based on a small time span for only one age. Evidence of construct validity would be appropriate. The test is normed only on "normal" children, although it is intended for use with exceptional children. Otherwise, care was taken in assembling a nationally representative sample.

VII. REFERENCES:

Buros, Oscar K. (ed.), The Eighth Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1978.

McCarthy, Dorothea, McCarthy Scales of Children's Abilities. Complete Test Kit. New York: The Psychological Corporation, 1972.

Review Form

TITLE: Metropolitan Readiness Tests

FORMS: Levels I and II, Forms P and Q
COPYRIGHT: 1974, 1976

AUTHORS: Joanne R. Nurss and Mary E. McGauvran

PUBLISHER: Harcourt Brace Jovanovich, Inc.

SOURCE: Harcourt Brace Jovanovich, Inc., 757 Third Avenue, New York, N.Y. 10017

PRICE AS OF 1979: $12.35 (package of 35 test booklets--each form within each level); $4.15 (specimen set)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: reading readiness test

INTENDED USE: to assess the "diverse range of pre-reading skills" in young children in order to plan instructional activities

INTENDED POPULATION: children in kindergarten and grade 1 (Level I for beginning and end of kindergarten; Level II for end of kindergarten and beginning of first grade)

FORMAT: multiple choice, with from four to six items per page; children mark answers by drawing a diagonal line from the upper right-hand corner to the lower left-hand corner, for machine scored tests

CONTENT:

	SUBTEST	SKILL AREA	NUMBER OF ITEMS	APPROXIMATE TIME	LEVEL
PRE-READING COMPOSITE	Auditory Memory	--------	12	15	I
	Rhyming		13	16	
	Letter Recognition	Visual	11	9	I
	Visual Matching		14	14	
	School Language and Listening	Language	15	14	I
	Quantitative Language		11	12	
PRE-READING COMPOSITE	Beginning Consonants	Auditory	13	25	II
	Sound-Letter Correspondence		16		
	Visual Matching	Visual	10	25	II
	Finding Patterns		16		II
	School Language	Language	9	15	II
	Listening		9		

(continued)

SUBTEST	SKILL AREA	NUMBER OF ITEMS	APPROXIMATE TIME	LEVEL
Quantitative Concepts	Quantitative	9	25	II
Quantitative Operations	(optional)	15		

An optional Copying Test is also available for both levels.

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST

There have been three earlier editions of the MRT: 1933 (developed by Dr. Gertrude H. Hildreth), 1949, and 1964. The basic purpose of all of the versions is "the assessment of readiness to begin formal school learning." According to the manual (part II), the 1976 edition "was developed to provide measures of some important pre-reading skills that are currently emphasized in kindergarten and Grade 1 instructional programs." The manual does not discuss the ways in which the newest edition differs from the previous editions.

Item analysis was conducted in a two-phase program in 1972 and 1973. In 1972, items were tested on 10,320 students in 40 school districts "to study the feasibility of measuring certain pre-reading skills not previously tested and to try out new types of items." As a result of this program, 955 items in 14 content areas were arranged in 17 experimental booklets, plus a practice booklet. The items were designed to be appropriate to all kindergarten and first grade students "regardless of their sex, ethnic background, urban/rural residence, socioeconomic status, or geographic region." These were tested in the second program, in 1973, on 17,190 pupils in 92 school districts (10,616 from 54 districts in the item-analysis study and 6,574 from 38 districts in special research studies). As a result, six tests were constructed for Level I and eight for Level II for final standardization based on statistical and practical (administrative ease, time, etc.) criteria. Items with "obvious defects" were eliminated, along with those showing ethnic bias (as judged by statistical procedures and minority group consultants' recommendations). As a further means of preventing ethnic bias, the authors determined that letter sounds being tested were equally common in Spanish and English, and they used pictures of different racial and ethnic group members. Finally, parallel forms were matched on content, item difficulty and item discrimination, with an average difficulty level of 65-70% for the total test.

III. PRACTICAL CONSIDERATIONS

MATERIALS REQUIRED: all provided in the test kit (test booklets, practice booklet, manual Parts I and II, key for handscoring test booklet, and class record)

TYPE OF ADMINISTRATION (group or individual): group

ADMINISTRATION TIME: Level I tests range from approximately 9 to 16 minutes, and are to be administered in seven sittings (including

the practice test). Level II tests are about 15 to 25 minutes in length (each), and are administered in five sittings.

SETTING: regular classroom; students should be tested in groups of not more than 15.

DIRECTIONS: the Teacher's Manual Part I: General Directions for Administering presents instructions to the teacher in red or green ink (Level I and II, respectively), while material to be read to the students in printed in black

SAMPLE QUESTIONS: practice booklets are included in both levels, to be administered a day or more in advance of the first test to allow for additional practice if needed

SCORING PROCEDURES: test booklets may be scored by hand or by machine

LANGUAGE OF ADMINISTRATION: English

TRAINING NEEDED TO ADMINISTER: no special training required

IV. TECHNICAL INFORMATION

NORMING POPULATION: 49,618 participants in Level I standardization and 40,259 in Level II from throughout the nation. In addition, 19,379 and 15,995 participants were included in auxiliary research for Levels I and II, respectively (equating forms and levels, large-city norms, Grade 1 norms, and validation).

NORMING SAMPLE METHOD: stratified random sample of U.S. school districts with 30 strata: six levels of enrollment and five of socioeconomic status, defined by median years of schooling of the school district population 25 years or older. Tables showing the breakdown of participants by sex, geographic region, population group, and prior pre-kindergarten or kindergarten experience are presented in the Manual Part II. Additional strata were used for parochial schools within each geographic region, public schools with enrollments of less than 300, and large-city schools of more than 100,000. Because the sampling unit was the school district, the data were weighted so that the number of students in each stratum was proportional to the number in the U.S. population in similar school systems. The April testing for Level I indicated that that level was too easy for Spring use, so interpolated mid-year norms for kindergarten are presented. Norms for Level I end-of-year (not recommended except in unusual cases) and for the Level II Battery Composite are not included but are available from the publisher. For equating tests school systems were chosen to represent a variety of geographic regions, community sizes, and socioeconomic levels. The Grade 1 study was based on a sample chosen to represent the population for whom Level I was expected to be most appropriate. There are no norms for the optional copying test or the Early School Inventory.

NORMING MIDPOINTS: November and April for Level I and September (Grade 1) and April (kindergarten) for Level II

SCALES AND SCORES: raw scores for the composite Pre-Reading Score can be converted to percentiles. Skill area and the individual test results can also be translated into stanines or "performance ratings" (verbal descriptions: High = stanines 7-9; Average = stanines 4-6; Low = stanines 1-3). In addition, scaled scores are available for the Pre-Reading Composite scores and the skill area raw scores.

VALIDITY: the Manual Part II discusses content and predictive validity. It says content validity was examined through an analysis of the beginning reading process and an extensive literature review to determine skills related to reading. The items measuring these skills were then examined during the item analysis. Predictive validity is relevant to the claim that MRT measures some skills necessary for success in early school learning. For Level I the Manual Part II presents correlations between the MRT, given in the fall, and the Metropolitan Achievement Tests (1970 edition, Primer, Form H), administered in the spring of the same year. The correlations for the subtests range from .41 to .68, and for the Composites from .68 to .72. For Level II, correlations are given with spring administration of the MAT (1970, Primary I battery, Form H) and the Stanford Achievement Tests (1973, Primary Level I battery, Form B). With the MAT, the correlations (reading and math) range from .54 to .71, with .70 for the Pre-Reading Composite and .72 to .73 for the Battery Composite. With the SAT, the range for all tests was .48 to .71, with .65 to .76 for the Pre-Reading Composite and .69 to .79 for the Battery Composite. No evidence regarding construct or concurrent validity is presented.

RELIABILITY: split-half, and Kuder-Richardson formula 20 reliabilities and means and standard deviations are given by test and composite for each form of both levels. Alternate form reliabilities and standard error of measurement, means, and standard deviations are also presented. All the Level I estimates range from .58 to .92 for the subtest and .85 to .95 for the Pre-Reading Composite. For Level II, the ranges are .68 to .93 for the subtests, .90 to .94 for the Pre-Reading Composite, and .92 to .95 for the Battery Composite.

V. OUTSIDE REVIEWS

Reviews were located only for the 1964 edition of the MRT. The Seventh Mental Measurements Yearbook contains reviews by Robert Dykstra (pp. 1175-1176) and Harry Singer (pp. 1176-1177). The 1976 edition contains substantial revisions, many in areas commented upon in the reviews.

Dykstra'a review was favorable. In particular, he commends the extensive discussion on interpreting test results, and notes the authors' emphasis on total scores, rather than subtest scores. He also points out the letter (performance) ratings and expectancy tables, as aids to interpretation. The letter ratings have been modified for the 1976 edition, and expectancy tables are no longer provided. Finally, Dykstra applauds the discussions of validity (content, predictive and construct, the latter not mentioned in the 1976 manual), reliability (split-half and alternative form), and the standardization group.

Singer's review is more guarded. He welcomes the revisions of two subtests since the 1949 edition (the Sentence test became a Listening test; and an Alphabet test was added), but he suggests that examples of satisfactory or unsatisfactory drawings would help the scoring of the copying test. Singer points out that the 1964 edition contained first grade norms only, and he notes the wide range of variability in the predictive validity scores between achievement and each category of readiness scores. Finally, Singer claims the MRT is of limited usefulness for handicapped and bilingual children, and notes the lower reliability scores associated with low SES. He concludes that the MRT does not provide much "educationally useful information" that could not be obtained equally well from teacher judgment.

The CSE-ECRC evaluations rate the total battery fair in all categories. The subtests ranged from fair to good in measurement validity, examinee appropriateness, and administrative usability, and from poor to fair in normed technical excellence.

VI. COMMENTS

GENERAL INFORMATION: designed to assess pre-reading skills, the MRT does not require that children be able to read. It does, however, depend on children's listening skills, even for those skill areas not specifically focused on auditory abilities. In this light, it is important to note that the directions state that test items are not to be repeated, unless unusual circumstances may have prevented the items from being heard.

THEORY, CONSTRUCTION, AND DEVELOPMENT: the 1976 edition was developed to test pre-reading skills not previously measured. The skill areas are said to reflect a literature search on the process of reading, although some areas are not directly tested that could arguably be important indicators of pre-reading skills (for example, auditory blending).

The MRT is exemplary in that it presents appropriate cautions about undue reliance on individual test scores, recommends obtaining additional information, and suggests specific sources of such additional information. In addition, it presents a guide to the kinds of questions and thought processes that should play a role in interpreting the scores.

PRACTICAL CONSIDERATIONS: the MRT requires few materials, and is said to require no special training to administer. Test procedures are clear and, for the most part, quite precise, which is appropriate to insure standardized administration. However, the instructions do not mention when in the school year it is appropriate to test, although the norms tables suggest a rather wide range of dates before and after the standardization dates during which tests could be administered and norm interpretations used.

Norms tables are presented in a manner that is easy to read. Unfortunately, though, some of the norms that users might need are not included among the technical information, but must be sent for separately.

Finally, the procedures for marking answers, particularly those for the machine scorable version, are fairly rigid. Answers must be marked in a rather limited manner or they will be incorrectly scored, and stray marks are a problem. In both hand- and machine-scorable versions, changing answers is rather cumbersome. Erasures are forbidden: the "wrong" answer is completely covered over with horizontal lines which obscures this choice, and the "right" answer is then indicated. However, these correction procedures may create problems if children change their minds several times but finally settle on the original answer. The MRT is to be commended for providing practice booklets and instructions in test-taking skills.

The Manual acknowledges the difficulties in administering group tests with young children. Page- and row-finders are included to help focus children's attention, although as the manual points out, the row-finders may be awkward for left-handed children. In the event a class contains both left- and right-handed children, uniform procedures may require abandoning the row-finders altogether. In addition, efforts to reduce distractability may be hindered by the number of items (4-6) per page.

The tests do not rely on an ability to read, and pictures are clearly drawn and easy to see. In addition, the tests are reasonably short. However, their brevity reduces their reliability.

TECHNICAL QUALITY: norming procedures are clearly described. The standardization sample is large, and appropriate steps were taken to assure that it is nationally representative. However, since the standardization program included children with and without prior pre-kindergarten or kindergarten experience, it would have been useful to have separate norms for those groups. Finally, the rather broad norming period is a source of concern: it spans three months.

The manual presents appropriate cautions concerning the reliability data and the comparability of scores across subtests.

VII. REFERENCES

Buros, O.K. (ed.) Seventh Mental Measurements Yearbook. New Jersey: Gryphon Press, 1972.

Hoepfner, R., Stern, C., and Nummedal, S. (Eds.) CSE-ECRC Preschool/ Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1971.

Nurss, J.R., and McGauvran, M.E. Metropolitan Readiness Tests. New York: Harcourt Brace Jovanovich, Inc., 1976.

Teacher's Manual, Part I: Directions for Administering, Level I, Form P, 1976.

Teacher's Manual, Part II: Interpretation and Use of Test Results, Level I, 1976.

Answer Booklet, Level I, Form P, 1974.

Teacher's Manual, Part I: Directions for Administering, Level II, Form P, 1976.

Teacher's Manual, Part II: Interpretation and Use of Test Results, Level II, 1976.

Answer Booklet, Level II, Form P, 1974.

Early School Inventory, 1976.

Parent-Teacher Conference Report, Level I, 1976.

Parent-Teacher Conference Report, Level II, 1976.

Practice Booklet, Levels I & II, Forms P & Q

Review Form

TITLE: Peabody Picture Vocabulary Test

FORMS: A and B
COPYRIGHT: 1959, 1965

AUTHOR: Lloyd M. Dunn, Ph.D.

PUBLISHER: American Guidance Service, Inc.

SOURCE: American Guidance Service, Publishers' Building, Circle Pines, Minn. 55014

PRICE AS OF 1979: $15.00 (complete kit, regular edition)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: individual intelligence test (of "verbal intelligence")

INTENDED USE: PPVT is a test of "hearing vocabulary" designed to provide an estimate of a child's verbal intelligence, in order to predict school success. Manual says it has "wide utility as a clinical tool."

INTENDED POPULATION: ages 2 1/2 to 18, particularly useful with a variety of special needs children

ITEM FORMAT: multiple choice (child identifies one of four pictures that correspond to spoken word); requires neither oral responses nor the ability to read.

CONTENT: the test consists of 150 test plates, each with four pictures, depicting 300 words, mostly singular or collective nouns, with some gerunds and a few adjectives and adverbs. The number of questions varies with each individual; i.e., the test is administered until a ceiling of 6 consecutive incorrect responses is reached.

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST

The PPVT is designed to measure hearing vocabulary, and is assumed to estimate the subject's verbal intelligence. To support this inference, the Manual refers to numerous studies that have shown vocabulary is the best single predictor of school success. The vocabulary items were selected from an initial pool of 3,885 words found in Webster's New Collegiate Dictionary (1953, G. & C. Merriam). All were words whose meanings could be illustrated by line drawings.

The item analysis consisted of two stages. The first stage involved illustrating 2,055 words, testing them on 360 two through 18-year-old subjects, and determining difficulty level. Two hundred test plates, each illustrating four words, were developed for the second stage and tested on 750 subjects in counterbalanced order. From this procedure, the final 150 plates were selected. Words depicted on each plate had the same difficulty level; showed appropriate linear growth in terms of percent passing at successive age levels; showed no sex, regional or cultural differences; were mostly singular or collective nouns, with some gerunds, a few adjec-

tives and adverbs; and were up-to-date and non-technical.

III. PRACTICAL CONSIDERATIONS

MATERIALS REQUIRED: all included in test kit: series of plates, answer booklets, test manual

TYPE OF ADMINISTRATION (group or individual): individual

ADMINISTRATION TIME: untimed, but usually takes 10-15 minutes to administer

SETTING: "a quiet room away from others." Examiner is encouraged to use praise to motivate students.

DIRECTIONS: included in the manual

SAMPLE QUESTIONS: three sample plates included, and additional trial series are included in the manual

SCORING PROCEDURE: by hand, while the test is being administered, to establish the basal and ceiling levels

LANGUAGE OF ADMINISTRATION: English

TRAINING NEEDED TO ADMINISTER: no special training required (just the ability to pronounce the test words correctly)

EARLY CHILDHOOD CONSIDERATIONS: several features should help reduce distractability and maintain attention: the test is short, and it should have a high interest value for young children. Pictures are clear, and they depict subjects that should appeal to children, such as other young children and animals. In addition, the test is individually administered. Three practice items are included, to familiarize children who are not test-wise with the task.

IV TECHNICAL INFORMATION

NORMING POPULATION: 4,012 white children, ages 2-6 to 18, from Nashville, Tennessee

NORMING SAMPLE METHOD: a census was made of all preschool children in the geographical areas served by four elementary schools in the Nashville City Schools that had a composite normal probability curve of Kuhlman-Finch IQ scores for its grade three and above students. These preschoolers were tested individually in their homes on two occasions.

NORMING MIDPOINTS: norming data were obtained between April and June

SCALES AND SCORES: age norms, standard score norms, and percentile norms

VALIDITY: the PPVT manual discusses several types of validity and cites 33 research reports on this topic. The manual claims that the test's content validity--the fact that the test measures hearing vocabulary--is derived from the use of all of the words in Webster's New Collegiate Dictionary that could be represented by a picture. Construct validity (the inference that the test measures verbal intelligence or scholastic aptitude) is supported by a variety of studies of the measurement of intelligence that have found vocabulary to be the best single predictor of school success. The manual notes that the PPVT measures receptive vocabulary, while many of these studies have concerned expressive vocabulary, and further observes that because intellect has many facets, the PPVT is not a comprehensive measure of intellectual ability. Nevertheless, the manual maintains that, although the PPVT measures a restricted sample of behavior, it can yield a useful prediction of school success, particularly in areas related to verbal intelligence. Discussion of item validity focuses on linear growth rates in terms of percent passing at successive age levels, although the manual cautions that one should not infer that individual children's vocabulary increases in that manner.

"Congruent validity" is discussed in terms of both mental ages and IQ's with Binet (1937 and 1960) and of IQ's with Wechsler (WISC-F, V, P; W-B; and WAIS). For the 1937 Binet, correlations with mental age scores ranged from .60 to .87; with IQ's from .43 to .92. With the 1960 Binet, the correlations with mental age scores ranged from .82 to .86. Median correlations for the PPVT and Wechsler scales ranged from .39 to .67.* The manual points out the need for data on correlations of PPVT and Binet IQ scores by age levels for all intelligence levels. Evidence of concurrent validity with the STEP, CAT, WRAT, SAT, and MAT, as well as a few others, is presented in terms of PPVT correlations. These range from .04 to .91, with the median in the .50's. (The .04 figure may be an unusual case; the next lowest was .32.) And the manual notes that the PPVT correlates about equally well with tests of language arts, social studies, and mathematics achievement. Finally, discussion of validity stresses the importance of data on predictive validity for the PPVT. However, the information available is limited. Only two studies reported on predictive validity, both using MAT scores as a criterion, and they yielded coefficients that range from .22 to .43.

RELIABILITY: the manual presents data on alternate-form reliability and standard errors of measurement. The coefficients for the early childhood ages (2 1/2 through 6) ranged from .67 to .81. The manual also cites 11 studies reported between 1959 and 1964 that provide reliability information on the PPVT with a variety of populations, and notes generally comparable coefficients to those found for the standardization population. A table summarizing this information is included in the manual. Although some data are available on temporal stability, the manual questions

*Correlations with the adult Wechsler scales are not included here, but can be found on page 33 of the manual.

the concept and asserts that test scores can be expected to fluctuate after some time from subject to subject, because behavior will be modified by individual experiences.

V. OUTSIDE REVIEWS

Buros' Sixth Mental Measurements Yearbook contains two reviews of the PPVT, one by Howard B. Lyman (pp. 820-821) and the other by Ellen V. Piers (pp. 822-823). Lyman praises the test for its interest level and its simple administration and scoring procedures. In general, he calls it a usable test. However, he also questions several aspects of the test. First, Lyman suggests that the norms might be biased to an unknown extent because they were based on the entire samples of cases and not on the half (approximately) who took a particular form first. The reviewer also notes that interpreting norms based on such a limited population should be done with caution. Second, he criticizes the 6-month (for 5 years and under) and 12-month chronological classifications on the norms tables. Such large intervals produce large jumps in IQ scores between children who are only one month apart but fall into different intervals. Third, Lyman discounts some of the reliability coefficients because they cover a wide range in ages. Finally, he maintains that the test's validity is "largely unestablished," since the studies reported on in the manual were considered by the author as limited and preliminary. It should be noted, however, that that disclaimer regarding "limited and preliminary" validity evidence has been removed from the 1965 edition of the manual, which reports the results of all the studies, 33 in number, that came to the author's attention in the test's first five years of use.

Piers praises the PPVT for its interest level, sharp pictures, and clear manual. Her concern about the lack of IQ norms outside the 55 to 145 range has been corrected in the most recent edition. Piers questions the test's predictive and congruent validity because of "the author's assumption that recognition vocabulary tests are said to measure it." Here again, the new test manual acknowledges the complexity of intelligence and no longer makes that assumption. Piers' most serious concern is that untrained users of the test might make inappropriate judgments about intelligence from a basically limited test. Nevertheless, she rates the PPVT as "probably now the best of its kind."

Finally, the PPVT receives fair ratings in the CSE-ECRC test evaluations in terms of examinee appropriateness, administrative usability, and normed technical excellence, and a good rating in terms of measurement validity, for both kindergarten and prekindergarten.

VI. COMMENTS

THEORY, CONSTRUCTION, AND DEVELOPMENT: because the PPVT tests hearing vocabulary by means of the identification of pictures, it focuses on the child's ability to label items; and it does not examine the function of items, nor their qualities nor relationships

between or among them. It is also a test of receptive vocabulary, and not of expressive vocabulary. In sum, the PPVT tests a limited sample of behavior. The test manual acknowledges this, and cautions against interpreting scores as comprehensive measures of intellectual functioning. While the manual claims the test may be useful as a clinical tool by providing a useful predictive measure of school success, it does not discuss more specific uses. It might have applications as a selection instrument, for example, but whether, or under what conditions, such a use would be appropriate, is not mentioned.

A noteworthy feature of the PPVT is that the test can apparently be used with handicapped children. This instrument is also unusual in that in constructing it, the author paid attention to the appeal of different pictures for people of different ages. Thus, earlier pictures are of young children, middle ones of elementary children, while later pictures are of youth and adults. Unfortunately, however, the illustrations are all of Caucasians--except for one black man, shown as a porter. In addition, all the people are portrayed in very traditional sex roles: females do housework, care for children, are teachers, dancers, or are in other traditionally "female" positions; males are physically active, and hold a large variety of occupations, such as truck driver, policeman, dentist, surveyor, judge, archaelogist, manual laborer, etc. The test could be improved considerably if the otherwise clearly drawn pictures were updated.

The test also suffers from a number of ambiguous items. In some instances, a singular word is depicted in the plural in picture form. In other cases, more than one picture in a given plate could justifiably be chosen as the correct answer. For example, in Form A, item 58 is the word "ceremony." The pictures show (1) either a person praying or a statue of someone apparently meditating,* (2) a man and a veiled woman carrying candles, (3) a waiter removing the lid from a serving bowl, and (4) two men shaking hands.

PRACTICAL CONSIDERATIONS: From a practical standpoint, the PPVT has much to recommend it. Directions and the pictures are clear, the materials are compact and easy to use, and administration requires no special preparation. The PPVT addresses many of the problems of testing young children. It is a short test, and one that should have a high interest value for children. Pictures are clear, and they depict "subjects" that should appeal to them, such as other children and animals. In addition, the test is individually administered. All of these features should help reduce distractability and maintain the child's attention. Three practice items

*Form B indicates that that picture is of an "idol."

are included in the book of plates, to be administered just before the test. The manual recommends that young children be given additional trials, if necessary, and provides vocabulary items for three more trial series. The examiner is also encouraged to help very immature children establish the correct pointing behavior by placing the child's finger on the correct answer during the first few trials. If the child does not learn the test-taking procedures, the examiner is advised to discontinue the test. A final note on the test's sensitivity to young children: the test includes a larger number of pictures for preschoolers than for older children. However, the manual contains no precautions about interpreting scores with young children, or about the instability of the scores.

TECHNICAL QUALITY: A commendable feature of the PPVT is its solicitation from researchers of evidence on the test's reliability, validity, or utility. However, some aspects of the test's technical character are questionable.

The first problem is the test's norms. Despite the procedures adopted to make the norms useful throughout the country, the norm group--all white children from Nashville, Tennessee--is not nationally representative. Clearly then, the test, and particularly the norms, should be used only with utmost caution with other types of children.

The discussion of alternate-form reliability and temporal stability is appropriate. However, the discussion of validity, though thorough, raises some important questions. Most notably, the primary use apparently intended for the PPVT is predictive, yet evidence of predictive validity is on the weak side. Only two studies examined it, and both found low correlations (from .22 to .43 with the MAT). Instead, what evidence is presented for predictive, concurrent, and congruent validity comes from 33 studies conducted during the first five years since the test's publication, and not directly from the author.

VII. REFERENCES

Buros, O.K. (Ed.) Sixth Mental Measurements Yearbook. New Jersey: Gryphon Press, 1965.

Dunn, L.M. Peabody Picture Vocabulary Test. Circle Pines, Minnesota: American Guidance Services, Inc., 1965.

Manual, 1965
Individual Test Record, Form a, 1959
Individual Test Record, Form b, 1959
Series of Plates, 1959

Hoepfner, R., Stern, C., & Nummedal, S. (Eds.) CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1971.

Review Form

TITLE: Stanford Early School Achievement Test

FORMS: One form only
COPYRIGHT: 1969, 1971

AUTHORS: Richard Madden and Eric F. Gardner

PUBLISHER: Harcourt Brace Jovanovich, Inc.

SOURCE: Harcourt Brace Jovanovich, Inc., 757 Third Ave., New York, NY 10017

PRICES AS OF 1978: $14.30 to $19.35 (package of 35 test booklets, directions, class record); $21.70 to $23.65 (package of 35 machine scoreable test booklets); $2.65 (specimen set, Level I); $2.95 (specimen set, Level II)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: group achievement test

INTENDED USE: to assess a child's background and mastery of experiences thus far in his or her school career; the SESAT is not intended as a readiness test per se

INTENDED POPULATION: kindergarten and first grade children (Level I for K.1 to 1.8; Level II for K.8 to 2.1)

ITEM FORMAT: multiple choice, 14 items per page

CONTENT:

Levels:	I		II	
Grades:	K.1-1.8		K.8-2.1	
Subtests	Items	Time	Items	Time
The Environment	42	25,15*	39	20,20*
Mathematics	28	15	55	20
Letters and Sounds	28	15	41	20
Aural Comprehension	28	20	27	20
Word Reading	--	--	58	20
Sentence Reading	--	--	39	20

*Test administered in two parts, including sample questions

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST:

This is the first edition of the SESAT. The impetus behind its construction is not stated in the teacher's manual.

The two levels of the test are similar in content. However, they differ in the type of knowledge on which they focus. According to the authors, Level I deals mainly with general experiential learning, the knowledge many children bring with them to school. Level II assumes that the children have had more formal instruction. The authors suggest that Level II be used to bridge the

gap between the first level of the SESAT and the Primary I level of the Stanford Achievement Test.

For both levels, the Environment subtest covers topics in the social and natural environment and sciences. Letters and Sounds deals with recognition of upper and lower case letters and beginning sounds. Ability to pay attention and to interpret and comprehend the spoken word is tested in Aural Comprehension. For Level I, Mathematics poses questions of conservation of number, space, and volume, counting, measurement, classification, and simple operations. Level II's Mathematics section is more formal and deals with addition, subtraction, the language of problem solving, and general mathematics concepts. In addition, at Level II, Word Reading and Sentence Reading subtests are given. The former calls for recognition of words selected from preprimers, primers, and first grade readers; sources are not listed. The latter includes questions of sentence structure, tenses, plurals, negation, and forms of "to be."

Individual items were chosen after an analysis of trial items administered to 3,100 first grade children in November, 1967, for Level I and to 2,004 first graders in November, 1968, for Level II. Item difficulties are reported in the Directions for Administering. No other information on content selection is provided.

III. PRACTICAL CONSIDERATION:

MATERIALS REQUIRED: all provided in test package (test booklets, teacher's manual, directions for administering)

TYPE OF ADMINISTRATION (group or individual): group

ADMINISTRATION TIME: total time takes from 1 1/2 to 2 1/3 hours, administered in five to seven sittings; individual subtests take from 15 to 25 minutes.

SETTING: classroom; "relaxed atmosphere" recommended, but children should be seated so as to prevent copying

DIRECTIONS: instructions for every item except those in Sentence Reading are read aloud by the teacher. The script is printed in boldface in the Directions for Administering. No instructions are included in the children's booklets.

SAMPLE QUESTIONS: 10-14 practice questions, administered as part of the first test sitting; no refresher practice questions in Level I, but some in Level II.

SCORING PROCEDURES: hand- and machine-scoring

LANGUAGE OF ADMINISTRATION: English and Spanish (Chicano and Puerto Rican editions) directions are available

TRAINING NEEDED TO ADMINISTER: no special training required

IV. TECHNICAL INFORMATION

NORMING POPULATION: Level I: 8,310 kindergarteners and 11,106 first graders. Level II: 7,369 first graders (October), 3,003 (Jan. - February), and 3,143 (April - May)

NORMING SAMPLE METHOD: exact method not specified, but the manual says the sample was chosen to reflect 1960 U.S. Census data in regard to geographic region, city size, and socioeconomic status in terms of median family income and median years of school completed by adults 25 years or older.

NORMING MIDPOINTS: Level I, beginning (October), and end of kindergarten. Level II, beginning (October), middle (end of January), and end (end of April) of Grade 1.

SCALES AND SCORES: raw scores for the subsections and the total test can be converted to percentiles and stanines

VALIDITY: the teacher's manuals do not mention validity. Correlations between the SESAT and the Otis-Lennon Mental Ability Test for Grade 1 are reported but their significance is not discussed.

RELIABILITY: Spearman-Brown Prophecy Formula split-half reliability coefficients, standard errors of measurement, and raw score standard deviations are presented by grade for Level I, and Kuder-Richardson Formula 20 reliability coefficients, raw score standard errors, and standard deviations by grade are given for Level II. These internal consistency measures range from .76 to .89 for Level I and from .612 to .970 for Level II. The authors note in each case that the low reliability coefficients are to be expected because the test is intentionally short.

V. OUTSIDE REVIEWS

The Stanford Early School Achievement Test is reviewed by Elizabeth Hagen and William A. Mehrens in the Seventh Mental Measurements Yearbook and by Courtney B. Cazden in the Eighth Mental Measurements Yearbook. These reviewers are in agreement on some aspects of the test series; their opinions differ on other facets.

One striking area of agreement is in regard to the lack of test validity for specific uses. Hagen, Mehrens, and Cazden all comment on the absence of evidence to support suggested test interpretations. All three reviewers also criticize the lack of data about the item analysis and other procedures used to choose the content of the tests. Finally, Hagen and Mehrens commend the SESAT for its clearly written directions.

There is disagreement among the three reviewers about the quality of the items. Mehrens writes, "In general, the questions seem to be well-written. They appear to have both face and content validity" (p. 55, 7th

MMY). Cazden, however, notes "The most serious problem with SESAT is the unknown or at best questionable validity of the test content" (p. 107, 8th MMY). She questions the content validity of the Environment subtests, explaining that:

> The environmental experience of young children in different parts of the country and in different cultural groups varies widely, and there is no common kindergarten social studies or science curriculum to lessen these differences . . . there seems no way at present to rationalize any particular 42 bits of information as either essential or equally representative for all. SESAT authors offer no such rationalization, and yet these 42 items make up on-third of the entire test.
>
> (p. 108, 8th MMY)

Another area of contention is the format. While Mehrens finds it to be very good, Cazden feels that the pages of the children's booklets contain too many items and pictures; children may be distracted or they may find it difficult to follow along.

Cazden states that "especially commendably, neither mental age nor grade-equivalent scores are given. As the revised [APA, AERA, and NCME] Standards urge, the use of such grossly misinterpreted scores should be abandoned or discouraged in all testing" (p. 108). Despite this convention the SESAT publisher does provide grade equivalent scores for the SESAT in Stanford Research Report #9.

Finally, the reviewers differ in their conclusions about the test. On the positive side, Hagen writes, "The SESAT is a well-constructed test with moderate reliability; it can be used by schools to assess their kindergarten program if the objectives and content of their program are similar to those appraised by the test" (p. 54). Mehrens concludes, "The SESAT is a needed type of instrument. Data from such an instrument should be potentially useful in instructing students" (p. 55). Cazden, however, states emphatically that "the lack of adequate information about content validity and the inferior test format make it impossible for the reviewer to give any recommendation for its use" (p. 108).

In terms of the CSE evaluations, both Levels I and II are consistently rated "poor" for individual subtests and the total test in the areas of measurement validity and normed technical excellence. However, both fare better on ratings on examinee appropriateness and administrative usability, with Level I designated "fair" on both and Level II "good."*

The SESAT is reviewed in Wahlstrom, Danley, and Raphael (1977), in terms of the curriculum objectives presented in the Ontario Ministry of Education publication, The Formative Years. A bibliography of other literature through 1977 concerning this instrument is included in the 8th MMY.

VI. COMMENTS:

DESCRIPTIVE INFORMATION: the SESAT provides two test levels, for children with and without formal kindergarten instruction. A more complete description of these terms would be helpful.

THEORY, CONSTRUCTION, AND DEVELOPMENT: very little information is given regarding construction of the tests. Methods of sampling are not stated. Although the Level I test is intended for kindergarten as well as first grade, the item analysis sample consisted almost entirely of first graders, except that "certain items that were thought to be oriented to the preschool child were placed in a separate booklet and administered to five- and six-year-old children to ascertain performance at a preschool level" (p. 19, Level I, Directions for Administering).

The grade appropriateness of some items in the Environment subtests is questionable. For instance, at Level I, children are asked to decide whether a wheel, a pail, or a rocking chair is made to move things more easily" (Question 30). Unless a child knows the abstract concept of a wheel as opposed to wheelbarrow, she or he might choose the pail, which is instantly capable of transporting water easily; however, wheel is intended as the correct answer. In other Level I questions, children are asked to recognize that electricity will not flow without a complete path and that a rock would be heavier than a watermelon of equal volume. At Level II, children are asked to recognize a line drawing of George Washington, to understand that "a judge serves all people equally," to recognize that whales are bigger than elephants, and to realize that "oxygen is essential for combustion" (p. 30, Directions for Administering).*

No empirical evidence is offered in support of the proposed test uses. Examples of test interpretation tend to be shallow at times:

> Linda V. is especially weak in Mathematics. She should be encouraged to manipulate blocks and other counters. She should be talked to in terms of quantity. Small books with mathematics stories might be sent home to be read to her. A good conceptual background should be built as a foundation for counting and computation. Linda should be taught to compare, match, and estimate. (p. 14, Level I, Directions for Administering)

At other times, conclusions are suggested for which no validity evidence is offered:

> Pupil B is somewhat below average in four of the six parts of the test. Such consistency may indicate a middle-of-the-class performer, now and in the years ahead. (p. 18, Level II, Directions for Administering)

Nor is there a discussion of children's varying developmental rates.

---

*In response to our comments on these items in a draft review of the SESAT, S. Miller of Harcourt Brace Jovanovich, Inc. points out, however, that "with one exception, these items have p-values of at least 50 and in some cases 60 or 70. Thus, at least 50 percent of the children answered these questions correctly" (Note 3).

PRACTICAL CONSIDERATIONS: test administration, seating, and timing are flexible, as they must be for young children. However, the effect of such variability on the use of standardized norms is not discussed. Nor, for that matter, are the standardization testing conditions described. Instructions in the teacher's manuals are clearly written.

A problem with the SESAT is its lack of attention to some of the special needs of young children in the context of testing. First, the test includes 14 items per page, a number likely to confuse some children. Second, although the test includes 10-14 practice questions, these are administered as part of the first test sitting and may not provide enough of an opportunity for children to become comfortable with test-taking procedures. None of the subsequent sittings in Level I and only some in Level II have any refresher practice items, and those in Level II seem to be administered as though they were part of the actual test questions. A third concern with the SESAT from an early childhood perspective is the size and detail of the pictures. The illustrations are small, some of the choices are drawn close together, and some are not easily distinguished visually. Moreover, some of the drawings contain lines and markings intended to denote a moving object or a glowing light, but these representations may be unfamiliar to a young child.

TECHNICAL INFORMATION: the method of sampling for the standardization program is not mentioned. Other information is missing, as well. Statistics on sample and national characteristics are given for Level I only, not for Level II. Although the two levels are differentiated by the extent of previous schooling expected of the test-takers, no data on kindergarten attendance or type of instructional program is given for the standardization sample.

The Directions for Administering does present a clear discussion of score variability.

There is no mention of validity. Only internal consistency is considered by the test authors. Alternate-form reliability is not discussed since there is only one form. Test-retest stability is not mentioned.

VII. REFERENCES:

REFERENCE NOTES:

Miller, S., Memorandum to H.F. Bligh, November 22, 1978.

REFERENCES:

Buros, Oscar K., Seventh Mental Measurements Yearbook. New Jersey: Gryphon Press, 1972.

Buros, Oscar K., Eighth Mental Measurements Yearbook. New Jersey: Gryphon Press, 1978.

Hoepfner, Ralph; Carolyn Stern; Susan G. Nummedal, et al. CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: UCLA Graduate School of Education, 1971.

Madden, Richard and Eric F. Gardner. Stanford Early School Achievement Test. New York: Harcourt Brace Jovanovich, Inc., 1969, 1970, 1971.

Directions for Administering, Level I, 1969.
Directions for Administering, Level II, 1971.
Stanford Early School Achievement Test, Level I, 1969.
Stanford Early School Achievement Test, Level II, 1970.

Wahlstrom, Merlin W.; Danley, Raymond R.; and Raphael, Dennis. Measuring Achievement at the Primary and Junior Levels. Toronto: Ministry of Education, Ontario, 1977.

Review Form

TITLE: Tests of Basic Experiences 2 FORMS: Level K and L
COPYRIGHT: 1978

AUTHOR: Margaret H. Moss, Ed.D.

PUBLISHER: CTB/McGraw-Hill, Inc.

SOURCE: CTB/McGraw-Hill, Inc., Del Monte Research Park, Monterey, CA 93940

PRICE AS OF 1979: $51.00 (complete battery, each level); $5.50 (specimen set)

I. DESCRIPTIVE INFORMATION

TYPE OF TEST: group achievement test

INTENDED USE: to measure some of the factors that contribute to readiness, specifically children's awareness of the world around them. More specifically, the manual notes two intended uses: (1) measuring program effectiveness and student progress, and (2) indicating strengths and weaknesses in an instructional program.

INTENDED POPULATION: preschool through first grade: Level K is for preschool through the start of grade one; Level L is for the end of kindergarten through the end of grade one

ITEM FORMAT: multiple choice, with one or two items per page

CONTENT: each level consists of the following:

Practice test	(6 items)	
Language test	(26 items)	General Concepts
Mathematics test	(26 items)	General Concepts
Science test	(26 items)	
Social Studies test	(26 items)	

In addition to individual subject area scores, a total score and a combined score for the language and mathematics tests can be calculated to serve as a General Concepts test.

II. THEORY, CONSTRUCTION, AND DEVELOPMENT OF TEST

TOBE 2 is a revised version of TOBE (copyright 1970, 1971). It is designed to assess children's awareness of the world around them. The test assumes that differences in their experiences and learning opportunities will affect children's success in school. Therefore, although TOBE 2 is not considered a readiness test per se, it is intended to measure "whether the student has had experiences that contribute to that preparedness (to learn)." The manual points out that other attributes such as attitude also contribute to readiness but that the TOBE 2 does not assess them.

The subtests in both levels test the same topics. The Mathematics subtest deals with order of numbers; counting; geometry; time and money;

weight, volume, and linear measurement; properties and operations; and fractions. Language items cover visual discrimination; alphabet knowledge, initial sounds, final sounds, rhyming, space and location; verb tense, sentence sense, context, comprehension, and reading terminology. Topics covered in the Science subtest are electricity and magnetism; force, motion, and mechanics; light, optics, and sound; chemistry; earth science and astronomy; animal reproduction and development; animal behavior and characteristics; and plant life. The Social Studies questions are based on geography and travel, environment and use of natural resources, human relations and behavior, occupations and the world of work, sociocultural geography, history and chronology, health and safety, and money and consumer behavior. The Examiner's Manual notes that the Mathematics and Language Tests cover topics most readily learned through some kind of formal instruction, while the Science and Social Studies tests deal with events children are likely to learn about in their daily lives. Thus, the manual concludes, the latter two tests are likely to provide "the purest measure of experiential richness."

The Examiner's Manual and the preliminary edition of the norms book contain no information on item selection or standardization, except to mention that the TOBE 2 evolved, in part, from two earlier tests, Test of Basic Information and Items of Space and Location. The TOBE 2 Norms and Technical Data Book, not yet available, is scheduled to contain "pertinent technical data." TOBE and TOBE 2 have the same rationale; however, TOBE 2 is said to encompass changes in norms, level of difficulty, format, response method, practice test, number of tests, and number of items, in response to needs expressed by customers.

III. PRACTICAL CONSIDERATIONS

MATERIALS REQUIRED: all included in test kit: test booklets

TYPE OF ADMINISTRATION (group or individual): group

ADMINISTRATION TIME: 30 to 45 minutes for each test; 20 minutes for the practice test. To be administered within one uninterrupted week.

SETTING: regular classroom, with good lighting, sufficient space to work, and enough distance between students to prevent copying. As few as five or six or up to 20 students may be tested at once, depending on maturity.

DIRECTIONS: the Examiner's Manual for each level presents specific directions for the practice test and the individual subtests. The text to be read aloud to the students is preceded by boldface, "SAY," while information for the examiner only is in italics.

SAMPLE QUESTIONS: a practice test of six items is provided; the directions encourage administering it far enough in advance to allow for additional practice if needed.

SCORING PROCEDURE: the test booklets come in machine-scorable and hand-scorable formats

LANGUAGE OF ADMINISTRATION: English (Spanish directions are available)

TRAINING NEEDED TO ADMINISTER: no special training required

IV. TECHNICAL INFORMATION

NORMING POPULATION:*

NORMING SAMPLE METHOD:*

NORMING MIDPOINTS: three norming dates: October, January, April

SCALES AND SCORES: the TOBE 2 Norms Book, Preliminary Edition provides conversion tables for raw scores to percentiles, stanines, and NCE's and scale scores for both levels. Prekindergarten norms are available for spring testing only.

VALIDITY:*

RELIABILITY:*

V. OUTSIDE REVIEWS

Reviews are available only for the original version of TOBE. Courtney Cazden commented on it in the Seventh Mental Measurements Yearbook (pp. 33-34). Esther Diamond reviewed it in the Eighth Mental Measurements Yearbook (pp. 119-120), while selections from articles by Steven Thurber and Barton B. Proger also appear there (p. 120 and pp. 120-121, respectively). In addition, the Eighth Mental Measurements Yearbook contains separate reviews for three of the subtests: Mathematics (reviewed by Leroy G. Callahan, p. 443), Science (reviewed by Arlen R. Gullickson, pp. 1392-1393), and Language (reviewed by Stephen M. Koziol, Jr., p. 149). It should be noted that many of the areas that were criticized in TOBE have been changed in TOBE 2.

Cazden raised some questions about the content of some test items, and noted occasional lack of clarity in some pictures. However, she concluded that, given the problems of creating a group test of basic concepts for young children, "the design of the test and conditions of its administration are probably as good as can be obtained." The test items Cazden found problems with are no longer present in the revised version.

Diamond concurs that the TOBE, overall, is a "carefully constructed, well-researched early childhood battery." She praises its attention to problems with young children's attention span, distractibility, and unfamiliarity with a structured situation, but questions the author's rationale for classifying some of the test items into the four subject areas. Her primary criticism concerns the use of traditional sex role models in which girls are generally depicted as passive, while boys are active. TOBE 2 no longer presents people in these stereotyped roles and has dropped the generic "he" that Diamond criticized in the TOBE manual and instructions.

---

*Information not available yet. A TOBE 2 Norms and Technical Data Book is in preparation.

Thurber judges the TOBE more harshly, primarily in terms of its content validity, which he feels has not been adequately documented. In addition, he argues that the "practical efficacy" of the test has not been demonstrated. Thurber further suspects the TOBE has a substantial verbal component, which he suggests should be verified by means of a factor analysis.

In contrast, Proger's review commends the "relatively nonverbal nature of the items," and notes that the TOBE should be useful as a screening device for isolating general types of difficulties in classes. Thurber cautions against using it for individual diagnosis, however, because of the small number of items testing each topic within the four subject areas. He, too, questions the classification of some of the items, and he points out that the TOBE does not primarily emphasize experiences rather than information, as the author claims.

Callahan calls the Mathematics subtest a "useful assessment tool." His only criticism concerns what he calls the "meager" evidence regarding content validity. Gullickson's review of the Science subtest lauds the administrative procedures, but challenges the validity information and questions TOBE's use for prediction or diagnosis. Finally, Koziol points out a few "troublesome areas" in some of the Language items, but generally praises the test's convenient administrative and recording procedures, and its broad scope.

The CSE-ECRC evaluations generally parallel the reviews noted above. All of the subtests are rated good in terms of administrative usability, and fair in examinee appropriateness; nearly all receive poor ratings in regard to measurement validity, while most are rated fair in normed technical excellence (except for most of Level L, which is rated poor).

VI. COMMENTS

THEORY, CONSTRUCTION, AND DEVELOPMENT: according to the manual, TOBE 2 does not test children's readiness, but their preparedness to learn. However, exactly what is meant by this distinction is unclear. Much of the relevant background information is not presented in the Examiner's Manual or the Norms Book, Preliminary Edition.

One intended use--determining instructional needs of individuals--is open to question. The problem is the brevity of the test. The number of items on each topic covered in the four subject areas is not given, nor are the specific items categorized by topic. But, given the limited number of items (26) in the test as a whole, it seems reasonable to question whether the sample of each domain is too small for accurate measurement.

PRACTICAL CONSIDERATIONS: TOBE 2 is a practical test that requires few materials and does not depend on children's ability to read. The procedures for administering TOBE 2 are very clear in most respects, and appropriate cautions about testing young children and guidelines for unusual situations are presented. However, the order in which the tests are to be given is not established and appears to be left to the examiner's discretion. In addition,

the number of test booklets may make handling TOBE 2 somewhat awkward.

The test format, for the most part, is well designed, but it is not consistent in machine- and hand-scored versions. In Level K the machine-scoreable books have two items per page, whereas the hand-scoreable books have only one. (Level L format is uniform for both versions, with two items per page.) No explanation is given for this difference. And while the number of items in both levels is an improvement over the original TOBE (three items per page), one item per page would pose less distraction for young children. Also TOBE 2 presents items on both sides of the page, with alternative pages screened, which also presents a possible distraction for young children. Interestingly, the early version of TOBE received praise from several reviewers because it printed items on only one side of the page. The author offers no explanation for the change.

TOBE 2 is attentive to other problems of testing young children. To reduce the problem of young children's unfamiliarity with test-taking procedures, teachers are instructed to conduct a practice session and to teach the necessary skills, ideally the day before the tests are administered. The Examiner's Manual notes the general lack of reliability in test scores with young children, and cautions against too much reliance on one score. And the manual recommends flexibility in allowing imprecise answer marks.

However, TOBE 2 is a group test, and as such makes some of the problems of testing at the early childhood level more pronounced. Instructions for the use of proctors are aimed at minimizing these problems.

Though TOBE 2 does not require reading, it does depend heavily on children's listening skills, which are not ostensibly skills which the test aims to measure. Moreover, some of the items involve two or three statements, containing a series of facts, followed by the question. Such items may strain a young child's memory capacity. In addition, some pictures are not as clear as they should be, even for an adult to understand: they are somewhat small, and sometimes too detailed. Finally, each test takes from 30 to 45 minutes, which is on the long side for such young children.

TECHNICAL QUALITY: most of the pertinent technical information--standardization procedures, item selection, reliability, and validity--is not yet available. This, of course, severely limits the assessment of the quality and appropriate use of the test. However, the Norms Book, Preliminary Edition presents a small amount of information which can be reviewed here.

The preliminary norms book presents norms for three testing times (October, January, and April) for kindergarten and first grade, but only April norms for prekindergarten, for each of the two test levels. It advises the user to be sure to use the table for the correct level and time of year, but it doesn't suggest a range of allowable testing times.

VII. REFERENCES

Buros, O.K. (Ed.) Seventh Mental Measurements Yearbook. New Jersey: Gryphon Press, 1972.

Buros, O.K. (Ed.) Eighth Mental Measurements Yearbook. New Jersey: Gryphon Press, 1978.

Hoepfner, R., Stern, C., & Nummedal, S. (Eds.) CSE-ECRC Preschool/ Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1971.

Moss, M.H. Tests of Basic Experiences 2. Monterey, CA: CTB/McGraw-Hill, 1978.

Examiner's Manual, Level K, 1978
Examiner's Manual, Level L, 1978
Norms Book, Preliminary Edition, 1978
TOBE 2 Practice Test, 1978
TOBE 2: Language, Level K, 1978
TOBE 2: Mathematics, Level K, 1978
TOBE 2: Science, Level K, 1978
TOBE 2: Social Studies, Level K, 1978
TOBE 2: Language, Level K (Machine Scorable), 1978
TOBE 2: Mathematics, Level K (Machine Scorable), 1978
TOBE 2: Language, Level L, 1978
TOBE 2: Mathematics, Level L, 1978
TOBE 2: Science, Level L, 1978
TOBE 2: Social Studies, Level L, 1978
TOBE 2: Language, Level L (Machine Scorable), 1978
TOBE 2: Mathematics, Level L (Machine Scorable), 1978

REFERENCES

Almy, M., & Genishi, C. Ways of studying children, an observation manual for early childhood teachers. New York: Teachers College Press, 1979.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974. Reprinted in O. Buros (Ed.), Tests in print. Highland Park, N.J.: Gryphon Press, 1974.

Anastasi, A. Psychological testing (4th edition). New York: MacMillan, 1976.

Anderson, S., et al. Priorities and directions for research and development related to measurement of young children: Report on task 2. Princeton, N.J.: Educational Testing Service, October 1972.

Arrington, R. Interrelations of the behavior of young children. New York: Columbia University Press, 1932 (Child Development Monograph No. 8).

Beck, M., & Stetz, F. Teachers' opinions of standardized test use and usefulness. Paper presented to the American Educational Research Association, San Francisco, CA, 1979.

Berk, R. Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins, 1980.

Bersoff, D., & Ericson, C. A precise and valid measure of behavior and behavior change in the classroom. Journal of School Psychology, 1972, 10 (4), 361-366.

Bloom, B. Stability and change in human characteristics. New York: John Wiley and Sons, 1964.

Boehm, A., & Weinberg, R. The classroom observer: A guide for developing observation skills. New York: Teachers College Press, 1977.

Bonjeen, C., Hill, R., & McLenmore, S.D. Sociological measurements: An inventory of scales and indices. San Francisco, CA.: Chandler Publishing Company, 1967.

Borich, G., et al. Convergent and discriminant validity of five classroom observation systems: Testing a model. Unpublished paper. University of Texas at Austin, 1977.

Borich, G., & Madden, S. Evaluating classroom instruction: A sourcebook of instruments. Reading, MA: Addison-Wesley, 1977.

Boyer, E., Simon, A., & Karafin, G. (Eds.) Measures of maturation: An anthology of early childhood observation instruments. Philadelphia, PA.: Research for Better Schools, 1973.

Bradley, R., & Caldwell, B. Issues and procedures in testing young children. (ERIC/TM Report No. 37). Princeton, N.J.: Educational Testing Service, 1974.

Bromen, S., Nichols, P., & Kennedy, W. Preschool IQ: Prenatal and early developmental correlates. Hillsdale, N.J.: Erlbaum, 1975.

Brooks, J., & Weintraub, M. A history of infant intelligence testing. In M. Lewis (Ed.), Origins of intelligence infancy and early childhood. New York: Plenum Press, 1976.

Brown, J., Sitts, M., & Yarborough, J. ERIC: What it can do for you and how to use it. Stanford, CA.: ERIC Clearinghouse on Information Resources, 1975.

Bryk, A.S., Apling, R., & Mathews, R. Developing an evaluation system for early childhood ESEA Title I: A feasibility analysis (draft). Cambridge, MA.: The Huron Institute, 1978.

Buros, O. (Ed.). Tests in print I. Highland Park, N.J.: Gryphon Press, 1961.

Buros, O. (Ed.). Tests in print II. Highland Park, N.J.: Gryphon Press, 1974

Buros, O. (Ed.). Sixth mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1965.

Buros, O. (Ed.). Seventh mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1972.

Buros, O. (Ed.). Eighth mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1978.

Carini, P. Documentary processes. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Carini, P. The art of seeing and the visibility of the person. Grand Forks, N.D.: North Dakota Study Group on Evaluation, University of North Dakota, 1979.

Cartwright, C.A., & Cartwright, G.P. Developing observation skills. New York: McGraw-Hill, 1974.

Carver, R. Two dimensions of tests: Psychometric and edumetric. American Psychologist 1974, 29, 512-518.

Cazden, C.B. Evaluation of learning in preschool education: Early language development. In B.S. Bloom, J.T. Hastings, & G.F. Madaus (Eds.). Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Champion, R.J., & Sear, A.M. Questionnaire response rates: A methodological analysis. In R. Cochrane (Ed.), Advances in social research: A book of readings in research methods. London, England: Constable, 1973.

Cicourel, A. et al. Language use and school performance. New York: Academic Press, 1974.

Cronbach, L. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd edition). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. Essentials of psychological testing (3rd edition). New York: Harper and Row, 1970.

Cronbach, L.J., Gleser, G., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

CTB/McGraw Hill. Comprehensive tests of basic skills. (Level A form S). Monterey, CA.: Author, 1974.

Epps, E. Situational effects in testing. In L. Miller, The testing of black students. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Flaugher, R. The many definitions of test bias. American Psychologist, July 1978, 671-679.

Garner, W.R. Rating scales, discriminability and information transmission. The Psychological Review, 1960, 67, 343-352.

Gaffney, R., & Maguire, T. Use of optically scored test answer sheets with young children. Journal of Educational Measurement, Summer 1971, 8(2).

Gesell, A., Ilg, F., & Ames, L. Infant and child in the culture of today. New York: Harper and Row, 1974.

Goodwin, W., & Driscoll, L. Handbook for measurement and evaluation in early childhood evaluation. San Francisco, CA.: Jossey-Bass, 1980..

Guilford, J.P., et al. Ratings should be scrutinized. Educational and Psychological Measurement, 1962, 22, 439-447.

Hambleton, R., & Eignor, D. Guidelines for evaluating criterion referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R., et al. Criterion referenced testing and measurement: review of technical issues and developments. Review of Educational Research, 1978, 48(1), 1-47.

Haney, W. A technical history of the national Follow Through evaluation. Cambridge, MA: The Huron Institute, 1977. (Also issued by the U.S. Office of Education as The Follow Through evaluation, Vol. V of the Follow Through Planned Variation Experiment Series.)

Haney, W. Summary notes on the HEW conference on achievement tests, and the basic skills. Cambridge, MA: The Huron Institute National Consortium on Testing Project, 1978.

Haney, W., et al. ESEA Title I early childhood education. Review of literature on evaluation and instrumentation. (Interim Report) Cambridge, Ma.: The Huron Institute, 1978.

Haney, W. et al. Basic skills assessment in South Carolina: How progressive an approach to minimum competency testing? National Consortium on Testing Staff Circular, No. 8, Cambridge, MA: The Huron Institute, January 1981.

Haney, W., & Pennington, N. Reanalysis of Follow Through parent and teacher data from Spring 1975. Cambridge, MA: The Huron Institute, 1978.

Haney, W., Schodlatz, D., Walker, D., & Weiss, S. ESEA Title I early childhood education: Review of literature on evaluation and instrumentation (interim report). Cambridge, MA: The Huron Institute, June 1978.

Haney, W. & Scott, L. Talking with children about tests: A pilot study of test item ambiguity. National Consortium on Testing Staff Circular, No. 7, Cambridge, MA: The Huron Institute, August, 1980.

Hilliard, A. Standardized testing and African-Americans: Building assessor competence in systematic assessment. In S. White & R. Tyler (Eds.) Testing, teaching and learning. Washington, D.C.: National Institute of Education, 1979.

Hobbs, N. The futures of children: Categories, labels, and their consequences. Washington, D.C.: Jossey-Bass, 1975.

Hoepfner, R., Stern, C., & Nummedal, S. (Eds.) CSE-ECRC preschool/kindergarten test evaluations. Los Angeles, CA.: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1971.

Hoepfner, R., et al. CSE elementary school test evaluations. Los Angeles, CA.: Center for the Study of Evaluation and Early Childhood Research Center, UCLA, 1976.

Hutt, S., & Hutt, C. Direct observation and measurement of behavior. Springfield, Il.: Thomas, 1970.

Hunt, J. Intelligence and experience. New York. Ronald, 1961.

Illingsworth, R. The development of the infant and the young child, normal and abnormal. London: Churchill & Livingstone, 1972.

Johnson, H. Preschool test descriptions, test matrix and correlated test descriptions. Springfield, IL: Thomas, 1979.

Johnson, O. Tests and measurements in child development: Handbook II. San Francisco, CA: Jossey-Bass, 1976.

Johnson, O., & Bommarito, J. Tests and measurements in child development. San Francisco, CA.: Jossey-Bass, 1971.

Kamii, C.K. Evaluation of learning in preschool education: Socioemotional, perceptual-motor, and cognitive development. In B.S. Bloom, J.T. Hastings, and G.F. Madaus (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Kounin, J. Discipline and group management in classrooms. New York: Holt, Rinehart and Winston, 1970.

Larry P. v. Riles. No. C-71. 227 ORFP. United States District Court N.D. California. October 16, 1979, Federal Supplement.

Leinhart, G. Transition rooms: Promoting maturation or reducing education? Journal of Educational Psychology, 1980, 72(1), 55-61.

Madaus, G., et al. The sensitivity of measures of school effectiveness. Harvard Educational Review, May 1979, 49(2), 207-230.

McClung, M. Competency testing: Potential for discrimination. Clearinghouse Review, September 1977, 439-448.

Medley, D., & Mitzel, H. Measuring classroom behavior by systematic observation. In N. Gage (Ed.), Handbook of research on teaching. Chicago, Il: Rand McNally, 1963.

Mehan, H. Structuring school structure. Harvard Educational Review, 1978, 48(1), 32-64.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, October 1975.

Miller, L. The testing of black children. Englewood Cliffs, N.J.: Prentice Hall, 1974.

Millman, J. Criterion-referenced measurement. In W. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA.: McCutchan 1974.

National Institute of Education. ERIC: A Profile. Washington, D.C.: author. No date.

Oakland, T. (Ed.). Psychological and educational assessment of minority children, New York: Brunner/Mazel, 1977.

Padilla, A. Cultural considerations: Hispanic-American. In S. White and R. Tyler (Eds.), Testing, teaching and learning. Report of a conference on research on testing. Washington, D.C.: U.S. Dept. of Health, Education and Welfare, October 1979.

Petersen, N., & Novick, M. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-31.

Pikunas, J. Human development: An emergent science. New York: McGraw-Hill, 1976.

Popham, J. (Ed.). Evaluation in education. Berkeley, CA.: McCutchan, 1974.

Popham, W.J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice Hall, 1978.

Porter, A., et al. Practical significance in program evaluation. American Educational Research Journal. Fall 1978, 15(4).

Raizen, S., & Bobrow, S. Design for a national evaluation of social competence in Head Start children. Santa Monica, CA: Rand, 1974.

Ramseyer, G., & Cashen, V. The effect of practice sessions on the use of separate answer sheets by first and second graders. Journal of Educational Measurement, Fall 1971, 8(3), 177-181.

Rude, F.T. Readiness tests: Implications for early childhood education. The reading teacher. March, 1973, 76(6), 572-580.

Simmons, R. A library user's guide to ERIC. Stanford, CA.: ERIC Clearinghouse on Information Resources, Stanford Center for Research and Development in Teaching, 1975.

Stallings, J. Learning to look. A handbook on classroom observation and teaching models. Belmont, CA.: Wadsworth, 1977.

Stallings, J., Kaskowitz, D., et al. Follow Through classroom observation evaluation 1972-73. Menlo Park, CA.: Stanford Research Institute, August 1974 (ED 104 891).

Walker, C., et al. Criterion-referenced test evaluations. Los Angeles, CA.: Center for the Study of Evaluation, UCLA, forthcoming.

Walker, D. Socioemotional measures for preschool and kindergarten children. San Francisco, CA.: Jossey-Bass, 1973.

Walker, D., Bane, M., & Bryk, A. The quality of the Head Start Planned Variation data. Cambridge, Ma.: The Huron Institute, 1973.

White, S., & Tyler, R. (Eds.). Testing, teaching and learning. Report of a conference on research on testing. Washington, D.C.: U.S. Dept. of Health, Education and Welfare, October 1979.

White, S., et al. Federal programs for young children: Volume I, Goals and standards of public programs for children; Volume II, Review of evaluation data for federally sponsored projects for children; Volume III, Recommendations for federal program planning; and Summary review and recommendations. Washington, D.C.: U.S. Government Printing Office, 1973.

Wolfensberger, W. The principle of normalization in human services. Toronto: National Institute on Mental Retardation, 1972.

Wright, H.F. Observational child study. In P.H. Mussen (Ed.), Handbook on research methods in child development. New York: Wiley, 1960.

Wright, H.F. Recording and analyzing child behavior. New York: Harper and Row, 1967.